Graduate Texts in Mathematics

GTM

Francis Clarke

Functional Analysis, Calculus of Variations and Optimal Control

Springer

Graduate Texts in Mathematics 264

Graduate Texts in Mathematics

Graduate Texts in Mathematics bridge the gap between passive study and creative understanding, offering graduate-level introductions to advanced topics in mathematics. The volumes are carefully written as teaching aids and highlight characteristic features of the theory. Although these books are frequently used as textbooks in graduate courses, they are also suitable for individual study.

For further volumes:
www.springer.com/series/136

Francis Clarke

Functional Analysis, Calculus of Variations and Optimal Control

Springer

Francis Clarke
Institut Camille Jordan
Université Claude Bernard Lyon 1
Villeurbanne, France

ISSN 0072-5285 Graduate Texts in Mathematics
ISBN 978-1-4471-4819-7 ISBN 978-1-4471-4820-3 (eBook)
DOI 10.1007/978-1-4471-4820-3
Springer London Heidelberg New York Dordrecht

Library of Congress Control Number: 2013931980

Mathematics Subject Classification: 46-01, 49-01, 49K15, 49J52, 90C30

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

To France. To its country lanes and market towns, its cities and cafés, its mountains and Roman ruins. To its culture and history, its wine and food, its mathematicians and its fast trains. To French society. To France.

Preface

The famous old road from Vézelay in Burgundy to Compostela in Spain is a long one, and very few pilgrims walk the entire route. Yet every year there are those who follow some part of it. We do not expect that many readers of this book will accompany us step by step from the definition of a norm on page 3, all the way to an advanced form of the Pontryagin maximum principle in the final chapter, though we would welcome their company. In describing the itinerary, therefore, we shall make some suggestions for shorter excursions.

The book consists of four parts. The first of these is on functional analysis, the last on optimal control. It may appear that these are rather disparate topics. Yet they share the same lineage: functional analysis (Part I) was born to serve the calculus of variations (Part III), which in turn is the parent of optimal control (Part IV). Add to this observation the need for additional elements from optimization and nonsmooth analysis (Part II), and the logic of the four parts becomes clear. We proceed to comment on them in turn.

Part I: Functional analysis. The prerequisites are the standard first courses in real analysis, measure and integration, and general topology. It seems likely to us, then, that the reader's backpack already contains *some* functional analysis: Hilbert spaces, at least; perhaps more. But we must set off from somewhere, and we do not, strictly speaking, assume this. Thus, Part I serves as an *introduction* to functional analysis. It includes the essential milestones: operators, convex sets, separation, dual spaces, uniform boundedness, open mappings, weak topologies, reflexivity...

Our course on functional analysis leads to a destination, however, as does every worthwhile journey. For this reason, there is an emphasis on those elements which will be important later for optimization, for the calculus of variations, and for control (that is, for the rest of the book).

Thus, compactness, lower semicontinuity, and minimization are stressed. Convex functions are introduced early, together with directional derivatives, tangents, and normals. Minimization principles are emphasized. The relevance of the smoothness of the norm of a Banach space, and of subdifferentials, is explained. Integral functionals are studied in detail, as are measurable selections. Greater use of optimization is made, even in proving classical results. These topics manage to walk hand in hand quite amiably with the standard ones.

The reader to whom functional analysis is largely familiar territory will nonetheless find Part I useful as a guide to certain areas.

Part II: Optimization and nonsmooth analysis. The themes that we examine in optimization are strictly mathematical: existence, necessary conditions, sufficient conditions. The goal is certainly not to make the reader an expert in the field, in which modeling and numerical analysis have such an important place. So we are threading our way on a fairly narrow (if scenic) path. But some knowledge of the subject and its terminology, some familiarity with the multiplier rule, a good understanding of the deductive (versus the inductive) method, an appreciation of the role of convexity, are all important things to acquire for future purposes. Some students will not have this background, which is why it is supplied here.

Part II also contains a short course on nonsmooth analysis and geometry, together with closely related results on invariance of trajectories. These subjects are certainly worth a detour in their own right, and the exposition here is streamlined and innovative in some respects. But their inclusion in the text is also based on the fact that they provide essential infrastructure for later chapters.

Part III: Calculus of variations. This is meant to be a rather thorough look at the subject, from its inception to the present. In writing it, we have tried to show the reader not only the landmarks, but also the contours of this beautiful and ongoing chapter in mathematics. This is done in part by advancing in stages, along a path that reveals its history.

A notable feature in this landscape is the presence of recent advanced results on regularity and necessary conditions. In particular, we encounter a refined multiplier rule that is completely proved. We know of no textbook which has such a thing; certainly not with the improvements to be found here. Another important theme is the existence question, where we stress the need to master the direct method. This is made possible by the earlier groundwork in functional analysis. There are also substantial examples and exercises, involving such topics as viscosity solutions, nonsmooth Lagrangians, the logarithmic Sobolev inequality, and periodic trajectories.

Part IV: Optimal control. Control theory is a very active subject that regularly produces new kinds of mathematical challenges. We focus here upon optimality, a topic in which the central result is the Pontryagin maximum principle. This important theorem is viewed from several different angles, both classical and modern, so as to fully appreciate its scope. We demonstrate in particular that its extension to nonsmooth data not only unifies a variety of special cases mathematically, but is itself of intrinsic interest.

Our survey of optimal control does not neglect existence theory, without which the deductive approach cannot be applied. We also discuss Hamilton-Jacobi methods, relaxation, and regularity of optimal controls. The exercises stem in part from several fields of application: economics, finance, systems engineering, and resources. The final chapter contains general results on necessary conditions for differential inclusions. These theorems, which provide a direct route to the maximum princi-

ple and the multiplier rule, appear here for the first time in a text; they have been polished and refined for the occasion.

A full proof of a general maximum principle, or of a multiplier rule, has never been an easy thing; indeed, it has been famously hard. One may say that it has become more streamlined; it has certainly become more general, and more unified; but it has not become easy. Thus, there is a difficult stretch of road towards the end of Part IV; however, it leads to a fully self-contained text.

Intended users. The author has himself used the material of this book many times, for various courses at the first-year or second-year graduate level. Accordingly, the text has been planned with potential instructors in mind. The main question is whether to do in detail (most of) Part I, or just refer to it as needed for background material. The answer must depend on the experience and the walking speed of the audience, of course.

The author has given one-semester courses that did not stray from Part I. For some audiences, this could be viewed as a second course on functional analysis, since, as we have said, the text adopts a somewhat novel emphasis and choice of material relative to other introductions. The instructor must also decide on how much of Chapter 8 to cover (it's nothing but problems). If the circumstances of time and audience permit, one could tread much of Part I and still explore some chapters from Part II (as an introduction to optimization and nonsmooth analysis) or Part III (the calculus of variations). As an aid in doing so, Part I has been organized in such a way that certain material can be bypassed without losing one's way. We refer especially to Sections 4.3–4.4, 5.4, 6.2–6.4, and Sections 7.2–7.4 (or all of Chapter 7, if in fact the audience is familiar with Hilbert spaces).

Here is a specific example. To give a course on functional analysis and calculus of variations, one could choose to travel lightly and drop from Part I the material just mentioned. Then, Chapter 9 might be done (minus the last section, perhaps). Following this, one could skip ahead to the first three or four chapters of Part III; they constitute in themselves a viable introduction to the calculus of variations. (True, the proof of Tonelli's theorem in Chapter 16 uses unseen elements of Chapter 6, but we indicate a shortcut to a special case that circumvents this.)

For advanced students who are already competent in functional analysis, an entirely different path can be taken, in which one focuses entirely on the latter half of the book. Then Part I (and possibly Part II) can play the role of a convenient and strangely relevant appendix (one that happens to be at the front), to be consulted as needed. As regards the teaching of optimal control, we strongly recommend that it be preceded by the first three or four chapters of Part III, as well as Section 19.1 on verification functions.

In addition to whatever merits it may have as a text, we believe that the book has considerable value as a reference. This is particularly true in the calculus of variations and optimal control; its advanced results make it stand out in this respect. But its ac-

cessible presentation of certain other topics may perhaps be appreciated too (convex analysis, integral semicontinuity, measurable selections, nonsmooth analysis, and metric regularity, for example). We dare to hope that it will be of interest to both the mathematics and the control engineering communities for all of these reasons, as well as to certain related ones (such as operations research and economics).

A word about the exercises: there are hundreds. Some of them stand side by side with the text, for the reader to meet at an early stage. But additional exercises (many of them original in nature, and more difficult) lie waiting at the end of each part, in a separate chapter. Solutions (full, partial, or just hints) are given for quite a few of them, in the endnotes. A list of notation is given at the beginning of the index.

Acknowledgements.

I shall end this preface on a more personal note. My mathematical education has been founded on the excellent books and the outstanding teachers that I have met along the way. Among the latter, I have long wanted to give thanks to Thomas Higgins of the order of Christian Brothers; to Norbert Schlomiuk of l'Université de Montréal; to Basil Rattray and Michael Herschorn of McGill University; to Terry Rockafellar, Victor Klee, and Ernest Michael at the University of Washington. In the endnotes, I mention the books that have had the greatest influence on me.

It is a pleasure to acknowledge the unfailing support of l'Institut Camille Jordan and l'Université de Lyon over the years, and that of le Centre national de recherche scientifique (CNRS). I have also benefited from regularly teaching at l'Ecole normale supérieure de Lyon.

The book was written in large part during the ten years in which I held a chair at l'Institut universitaire de France; this was a major contribution to my work. Thanks are also due to Pierre Bousquet for his many insightful comments. And to Carlo Mariconda, who read through the entire manuscript (Vézelay to Compostela!) and made countless helpful suggestions, I say: *mille grazie*.

On quite a different plane, I avow my heartfelt gratitude to my wife, Gail Hart, for her constant love and companionship on the journey.

Francis Clarke
Burgundy, France

July 2012

Contents

Part I

Functional Analysis

Chapter 1

Normed Spaces

There are only two kinds of math books: those you cannot read beyond the first sentence, and those you cannot read beyond the first page.
C. N. Yang

What we hope ever to do with ease, we must learn first to do with diligence.
Samuel Johnson

We now set off on an expedition through the vast subject of functional analysis. No doubt the reader has some familiarity with this place, and will recognize some of the early landmarks of the journey. Our starting point is the study of *normed spaces*, which are situated at the confluence of two far-reaching mathematical abstractions: vector spaces, and topology.

1.1 Basic definitions

The setting is that of a vector space over the real numbers $\mathbb{R}$. There are a dozen or so axioms that define a vector space (the number depends on how they are phrased), bearing upon the existence and the properties of certain operations called addition and scalar multiplication. It is more than probable that the reader is fully aware of these, and we shall say no more on the matter. We turn instead to the central idea of this chapter.

A **norm** on the vector space X corresponds to a reasonable way to measure the *size* of an element, one that is consistent with the vector operations. Given a point $x \in X$, the norm of x is a nonnegative number, designated $\|x\|$. We also write $\|x\|_X$ at times, if there is a need to distinguish this norm from others. In order to be a norm, the mapping $x \mapsto \|x\|$ must possess the following properties:

- $\|x\| \geqslant 0 \;\; \forall x \in X$; $\|x\| = 0$ if and only if $x = 0$ (*positive definiteness*);
- $\|x+y\| \leqslant \|x\| + \|y\| \;\; \forall x, y \in X$ (*the triangle inequality*);
- $\|tx\| = |t|\,\|x\| \;\; \forall t \in \mathbb{R}, x \in X$ (*positive homogeneity*).

Once it has been equipped with a norm, the vector space X, or, more precisely perhaps, the pair $(X, \|\cdot\|)$, is referred to as a **normed space**.

F. Clarke, *Functional Analysis, Calculus of Variations and Optimal Control*, Graduate Texts in Mathematics 264, DOI 10.1007/978-1-4471-4820-3_1,

We have implied that vector spaces and topology are to meet in this chapter; where, then, is the topology? The answer lies in the fact that a norm induces a **metric** on X in a natural way: the distance d between x and y is $d(x,y) = \|x-y\|$. Thus, a normed space is endowed with a metric topology, one (and this is a crucial point) which is compatible with the vector space operations.

Some notation. The closed and open **balls** in X are (respectively) the sets of the form

$$B(x,r) = \{y \in X : \|y-x\| \leqslant r\}, \; B^{\circ}(x,r) = \{y \in X : \|y-x\| < r\},$$

where the radius r is a positive number. We sometimes write B or B_X for the closed unit ball $B(0,1)$, and B° for the open unit ball $B^{\circ}(0,1)$. A subset of X is **bounded** if there is a ball that contains it.

If A and C are subsets of X and t is a scalar (that is, an element of $\mathbb{R}$), the sets $A+C$ and tA are given by

$$A+C = \{a+c : a \in A, c \in C\}, \quad tA = \{ta : a \in A\}.$$

(Warning: $A+A$ is different from $2A$ in general.) Thus, we have $B(x,r) = \{x\} + rB$. We may even ask the reader to tolerate the notation $B(x,r) = x + rB$. The **closure** of A is denoted $\operatorname{cl} A$ or $\overline{A}$, while its **interior** is written $\operatorname{int} A$ or A°.

Given two points x and y in X, the **closed interval** (or segment) $[x,y]$ is defined as follows:

$$[x,y] = \{z = (1-t)x + ty : t \in [0,1]\}.$$

When t is restricted to $(0,1)$ in the definition, we obtain the open interval (x,y). The half-open intervals $[x,y)$ and $(x,y]$ are defined in the evident way, by allowing t to vary in $[0,1)$ and $(0,1]$ respectively.

The compatibility between the vector space and its norm topology manifests itself by the fact that if U is an open subset of X, then so is its translate $x+U$ and its scalar multiple tU (if $t \neq 0$). This follows from the fact that balls, which generate the underlying metric topology, cooperate most courteously with the operations of translation and dilation:

$$B(x,r) = x + B(0,r), \;\; B(0,tr) = tB(0,r) \;\; (t > 0).$$

It follows from this, for example, that we have $\operatorname{int}(tA) = t \operatorname{int} A$ when $t \neq 0$, and that a sequence x_i converges to a limit x if and only if the difference $x_i - x$ converges to 0. There are topologies on X that do not respect the vector operations in this way, but they are of no interest to us. We shall have good reasons later on to introduce certain topologies on X that *differ* from that of the norm; they too, however, will be compatible with the vector operations.

A vector space always admits a norm. To see why this is so, recall the well-known consequence of Zorn's lemma which asserts that any vector space has a basis $\{e_\alpha\}$, in the sense of linear algebra. This means that any x has a unique representation $x = \sum_\alpha x_\alpha e_\alpha$, where all but a finite number of the coefficients x_α are 0. The reader will verify without difficulty that $\|x\| := \sum_\alpha |x_\alpha|$ defines a norm on X. In practice, however, there arises the matter of choosing a *good* norm when a space presents multiple possibilities.

Sometimes a good norm (or any norm) is hard to find. An example of this: the space of all continuous functions f from $\mathbb{R}$ to $\mathbb{R}$. Finding an explicit norm on this space is problematic, and the space itself has found use only when endowed with a topology that is not that of a norm. In other cases, several norms may come to mind.

1.1 Example. The vector space of continuous functions f from $[0,1]$ to $\mathbb{R}$ admits, among others, the two following norms:

$$\|f\|_\infty = \max_{t \in [0,1]} |f(t)|, \quad \|f\|_1 = \int_0^1 |f(t)|\,dt,$$

both of which are well defined. One of these (the first, it turns out) is a better choice than the other, for reasons that will become completely clear later. □

Two norms $\|\cdot\|_1$ and $\|\cdot\|_2$ on X are said to be **equivalent** if there exist positive constants c, d such that

$$\|x\|_1 \leqslant c\|x\|_2, \quad \|x\|_2 \leqslant d\|x\|_1 \ \ \forall x \in X.$$

As the reader may verify, this amounts to saying that each ball around 0 of one type contains a ball around 0 of the other type. This property, in turn, is easily seen to characterize the equality of the topologies induced by the two metrics. Thus we may say: two norms on X are equivalent if and only if they induce the same metric topology.

1.2 Exercise. Are the two norms of Example 1.1 equivalent? □

When we restrict attention to a linear subspace of a normed space, the subspace is itself a normed space, since the restriction of the norm is a norm. This is an *internal* mechanism for creating smaller normed spaces. Another way to create new (larger) normed spaces is external, via Cartesian products. In this case, there is some flexibility in how to define the norm on the product.

Let X and Y be normed spaces, with norms $\|\cdot\|_X$ and $\|\cdot\|_Y$. Then the Cartesian product $Z = X \times Y$ may be equipped with any of the norms

$$\|x\|_X + \|y\|_Y, \quad \big\{(\|x\|_X)^2 + (\|y\|_Y)^2\big\}^{1/2}, \quad \max\big\{\|x\|_X, \|y\|_Y\big\},$$

among others. That these are all norms on $X \times Y$ is simple to check, and it is not much more work to go on to verify that all these **product norms** are equivalent.

1.3 Example. We denote Euclidean n-space by $\mathbb{R}^n$; this is the vector space consisting of n-tuples $x = (x_1, x_2, \ldots, x_n)$ of real numbers. By default, we always consider that it is equipped with the **Euclidean norm**

$$|x| = \left\{ x_1^2 + x_2^2 + \cdots + x_n^2 \right\}^{1/2}, \quad x \in \mathbb{R}^n.$$

No other norm is awarded the honor of being written with single bars. □

1.4 Example. (Continuous functions on a compact set) Let K be a compact metric space. We denote by $C(K)$, or $C(K,\mathbb{R})$ if more precision is desired, the vector space of continuous functions $f : K \to \mathbb{R}$, equipped with the norm

$$\|f\| = \|f\|_{C(K)} = \max_{x \in K} |f(x)|.$$

□

Notation: When K is an interval $[a,b]$ in $\mathbb{R}$, we write $C[a,b]$ for $C(K)$.

1.5 Exercise. Prove that $C[0,1]$ is an infinite dimensional vector space, by exhibiting a linearly independent set with infinitely many elements. □

1.6 Example. (Spaces of sequences) Some useful examples of normed spaces are obtained by considering sequences with certain properties. For fixed $p \in [1, \infty)$, we define

$$\|x\|_p = \left\{ \sum_{i \geqslant 1} |x_i|^p \right\}^{1/p},$$

where $x = (x_1, x_2, \ldots)$ is any sequence of real numbers. As we show below, the set of all sequences x for which $\|x\|_p < \infty$ is a vector space. It is equipped with the norm $\|\cdot\|_p$, and designated ℓ^p. The vector space of all *bounded* sequences, denoted ℓ^∞, is turned into a normed space with the norm $\|x\|_\infty := \sup_{i \geqslant 1} |x_i|$.

We shall require **Hölder's inequality**, which, in the present context, asserts that

$$\sum_{i \geqslant 1} |x_i y_i| \leqslant \|x\|_p \|y\|_{p_*}$$

for any two sequences $x = (x_1, x_2, \ldots)$ and $y = (y_1, y_2, \ldots)$. Here, p lies in $[1, \infty]$, and p_* signifies the **conjugate exponent** of p, the unique number in $[1, \infty]$ such that $1/p + 1/p_* = 1$. (Thus, $p_* = \infty$ when $p = 1$, and vice versa.)

Other members of the cast include the following spaces of convergent sequences:

$$c = \left\{ x = (x_1, x_2, \ldots) : \lim_{i \to \infty} x_i \text{ exists} \right\}, \quad c_0 = \left\{ x = (x_1, x_2, \ldots) : \lim_{i \to \infty} x_i = 0 \right\}.$$

Another vector space of interest consists of those sequences having finitely many nonzero terms; it is denoted ℓ_c^∞. (The letter c in this context stands for *compact*

support.) We equip c, c_0, and ℓ_c^∞ with the same norm as ℓ^∞, of which they are evidently linear subspaces. It is easy to see that all these normed spaces are infinite dimensional; that is, an algebraic basis for the underlying vector space must have infinitely many elements. □

1.7 Proposition. *For* $1 \leqslant p \leqslant \infty$, ℓ^p *is a vector space and* $\|\cdot\|_p$ *is a norm on* ℓ^p.

Proof. The cases $p = 1$, $p = \infty$ are simple, and are left to the reader as exercises. For $1 < p < \infty$, the inequality[1]

$$(a+b)^p \leqslant 2^p\big(a^p + b^p\big),\ \ a, b \geqslant 0$$

implies that ℓ^p is stable under addition. It is clearly stable under scalar multiplication as well; thus, ℓ^p is a vector space. We need only verify that the putative norm $\|\cdot\|_p$ satisfies the triangle inequality. To this end, we observe first the following:

$$|x_i + y_i|^p \leqslant |x_i + y_i|^{p-1}|x_i| + |x_i + y_i|^{p-1}|y_i|.$$

Then, we take the sum over i, and we proceed to invoke Hölder's inequality for the two terms on the right; the triangle inequality results. □

1.8 Exercise.

(a) Let $1 \leqslant p < q \leqslant \infty$. Show that $\ell^p \subset \ell^q$, and that the injection (that is, the identity map $\Lambda x = x$ from ℓ^p to ℓ^q) is continuous.

(b) It is clear that ℓ^1 is a subspace of c_0, and that c_0 is a subspace of ℓ^∞. In which case can we say *closed* subspace? □

1.9 Example. (Lebesgue spaces) We proceed to revisit some familiar facts and establish some notation. Let Ω be a nonempty open subset of $\mathbb{R}^n$, and let $f : \Omega \to \mathbb{R}$ be a Lebesgue measurable function. We write dx for Lebesgue measure on $\mathbb{R}^n$. The (Lebesgue) integral

$$\int_\Omega |f(x)|\,dx$$

is then well defined, possibly as $+\infty$. When it is finite, we say that f is *summable* (on Ω). The class of summable functions f is denoted by $L^1(\Omega)$. More generally, for any $p \in [1, \infty)$, we denote by $L^p(\Omega)$, or by $L^p(\Omega, \mathbb{R})$, the set of all functions f such that $|f|^p$ is summable, and we write

$$\|f\|_p = \|f\|_{L^p(\Omega)} := \left\{\int_\Omega |f(x)|^p\,dx\right\}^{1/p}.$$

There remains the case $p = +\infty$. The function f is *essentially bounded* when, for some number M, we have $|f(x)| \leqslant M$ a.e. The abbreviation "a.e." stands for "almost everywhere," which in this context means that the inequality holds except for

[1] This inequality will be evident to us quite soon, once we learn that the function $t \mapsto t^p$ is convex on the interval $(0, \infty)$; see page 36.

the points x in a null set (that is, a subset of Ω of measure zero). We define $L^\infty(\Omega)$ to be the class of measurable functions $f:\Omega\to\mathbb{R}$ that are essentially bounded, with norm

$$\|f\|_\infty = \|f\|_{L^\infty(\Omega)} = \inf\left\{ M : |f(x)| \leqslant M,\ x\in\Omega \text{ a.e.}\right\}.$$

Since the infimum over the empty set is $+\infty$, we see that $\|f\|_\infty = +\infty$ precisely when f fails to be essentially bounded. Thus, for any p in $[1,\infty]$, we may say that a measurable function f belongs to $L^p(\Omega)$ if and only if $\|f\|_p < \infty$. □

1.10 Exercise. Let $f\in L^\infty(\Omega)$. Prove that $|f(x)| \leqslant \|f\|_\infty,\ x\in\Omega$ a.e. □

Notation: When Ω is an interval (a,b) in $\mathbb{R}$, we write $L^p(a,b)$ for $L^p(\Omega)$.

We shall affirm below that $L^p(\Omega)$ is a vector space and $\|\cdot\|_p$ is a norm, but let us first recall a familiar convention. We identify two elements f and g of $L^p(\Omega)$ when $f(x)=g(x)$ for almost every $x\in\Omega$. Thus, the elements of $L^p(\Omega)$ are really equivalence classes of functions $\{f\}$, where $g\in\{f\}$ if and only if $f(x)=g(x)$ a.e. This distinction is not reflected in our notation,[2] but it explains why it is absurd to speak of (for example) the set of functions $f\in L^p(0,1)$ satisfying $f(1/2)=0$: this does not correspond to a property of an equivalence class. On the other hand, it makes sense to speak of those $f\in L^p(\Omega)$ which satisfy $|f(x)|\leqslant 1$ a.e. in Ω: this property is stable with respect to elements in an equivalence class.

Let us once again recall Hölder's inequality, which, in the current context, affirms that if $f\in L^p(\Omega)$ and $g\in L^{p_*}(\Omega)$, where $1\leqslant p\leqslant\infty$, then the function fg belongs to $L^1(\Omega)$, and we have

$$\|fg\|_1 \leqslant \|f\|_p\,\|g\|_{p_*}.$$

We can use this to adapt the proof of Prop. 1.7, and we obtain

1.11 Proposition. *For each $p\in[1,\infty]$, the class $L^p(\Omega)$ is a vector space, and $\|\cdot\|_p$ is a norm on $L^p(\Omega)$.*

1.12 Exercise. Show that $L^q(0,1)$ is a strict subspace of $L^p(0,1)$ if $1\leqslant p<q\leqslant\infty$. (This is true generally of $L^p(\Omega)$, when Ω is bounded; compare Exer. 1.8.) □

1.13 Example. (Absolutely continuous functions) Let $[a,b]$ be an interval in $\mathbb{R}$. A function $x:[a,b]\to\mathbb{R}$ is said to be *absolutely continuous* if x is continuous in the following sense: for every $\varepsilon>0$, there exists $\delta>0$ such that, for every finite collection $\{[a_i,b_i]\}$ of disjoint subintervals of $[a,b]$, we have

$$\sum_i (b_i-a_i) < \delta \implies \sum_i |x(b_i)-x(a_i)| < \varepsilon.$$

It is shown in elementary courses in integration that a continuous function x possesses this property if and only if it is an indefinite integral; that is, there exists a function $v\in L^1(a,b)$ such that

[2] Rudin: "We relegate this distinction to the status of a tacit understanding."

$$x(t) = x(a) + \int_a^t v(\tau)\, d\tau, \quad t \in [a,b]. \tag{$*$}$$

In this case, the theory of integration tells us that x is differentiable at almost every point in (a,b), with $x'(t) = (d/dt)x(t) = v(t)$, $t \in (a,b)$ a.e. Thus, absolutely continuous functions are well behaved, in that they coincide with the integral of their derivative. For this reason, they constitute the customary class in which the theory of ordinary differential equations is developed, and they will play a central role later when we study the calculus of variations.

The vector space of absolutely continuous functions on $[a,b]$ is denoted $\mathrm{AC}[a,b]$, and we endow it with the norm

$$\|x\|_{\mathrm{AC}} = |x(a)| + \int_a^b |x'(t)|\, dt\,.$$

More generally, for $1 \leqslant p \leqslant \infty$, we denote by $\mathrm{AC}^p[a,b]$ the class of continuous functions x which admit a representation of the form $(*)$ with $v \in L^p(a,b)$. The norm on $\mathrm{AC}^p[a,b]$ is given by

$$\|x\|_{\mathrm{AC}^p} = |x(a)| + \|x'\|_{L^p(a,b)}\,.$$

The function x on $[a,b]$ is called *Lipschitz* if there exists M such that

$$|x(s) - x(t)| \leqslant M|s-t| \;\; \forall s,t \in [a,b].$$

Such a function x is easily seen to be absolutely continuous, with a derivative x' (almost everywhere) that satisfies $|x'(t)| \leqslant M$ a.e. Thus, a Lipschitz function belongs to $\mathrm{AC}^\infty[a,b]$. Conversely, one shows that an element x of $\mathrm{AC}^\infty[a,b]$ satisfies the Lipschitz condition above, the minimal M for this being $\|x'\|_{L^\infty(a,b)}$. □

1.14 Exercise. Show that the function $x(t) = \sqrt{t}$ is absolutely continuous on the interval $[0,1]$, but is not Lipschitz. □

1.2 Linear mappings

A *linear* map (or application, or transformation) Λ between two vector spaces X and Y is one that exhibits a healthy respect for the underlying vector space structure; it preserves linear combinations:

$$\Lambda(t_1x_1 + t_2x_2) = t_1\Lambda(x_1) + t_2\Lambda(x_2), \quad x_1, x_2 \in X,\, t_1, t_2 \in \mathbb{R}\,.$$

Such maps turn out to be of central importance in the theory of normed spaces. Note that they constitute in themselves a vector space, since a linear combination of two linear maps is another linear map.

Notation: When Λ is linear, $\Lambda(x)$ is also written as Λx or $\langle \Lambda, x \rangle$.

The vector space of linear applications from X to Y is denoted by $L(X,Y)$. If Λ belongs to $L(X,Y)$, then $\Lambda(0) = 0$ necessarily. If Λ is continuous at 0, then $\Lambda^{-1}(B_Y)$ contains a neighborhood of 0, and therefore a ball rB_X in X. Thus

$$\|x\|_X \leqslant r \implies \Lambda x \in B_Y,$$

or, in a formulation that is easily seen to be equivalent to this,

$$\|\Lambda x\|_Y \leqslant (1/r)\|x\|_X \;\; \forall x \in X.$$

It also follows readily that the continuity of Λ at 0 is equivalent to its continuity everywhere, and to its being bounded above on a neighborhood of 0. In summary, and without further proof, we may say:

1.15 Proposition. *Let X and Y be normed spaces, and let $\Lambda \in L(X,Y)$. Then the following are equivalent:*

(a) *Λ is continuous;*

(b) *Λ is bounded above on a neighborhood of 0;*

(c) *There exists M such that $\|\Lambda x\|_Y \leqslant M \;\; \forall x \in B(0,1)$;*

(d) *There exists M such that $\|\Lambda x\|_Y \leqslant M\|x\|_X \;\; \forall x \in X$.*

1.16 Exercise. Let y_i $(i = 1,2,\ldots,n)$ be elements in a normed space Y, and let $\Gamma : \mathbb{R}^n \to Y$ be defined by

$$\Gamma\lambda = \Gamma(\lambda_1,\lambda_2,\ldots,\lambda_n) = \textstyle\sum_{i=1}^{n} \lambda_i y_i\,.$$

Prove that Γ is continuous. □

The elements of $L(X,Y)$ are often referred to as **operators**. We reserve the term *linear functional* for the elements of $L(X,\mathbb{R})$; that is, the real-valued linear applications on X. For any element Λ of $L(X,Y)$ we write

$$\|\Lambda\| = \|\Lambda\|_{L(X,Y)} = \sup\left\{ \|\Lambda x\|_Y : x \in X,\; \|x\|_X \leqslant 1 \right\}.$$

1.17 Exercise. Let $\Lambda \in L(X,Y)$, where X,Y are normed spaces. Then

$$\|\Lambda\| = \sup_{x \in X,\, \|x\|=1} \|\Lambda x\|_Y = \sup_{x \in X,\, \|x\|<1} \|\Lambda x\|_Y = \sup_{x \in X,\, x \neq 0} \frac{\|\Lambda x\|_Y}{\|x\|_X}.$$ □

The reader will notice that two of the expressions displayed in this exercise are inappropriate if X happens to be the trivial vector space $\{0\}$. Thus, it is implicitly assumed that the abstract space X under consideration is nontrivial; that is, that X contains nonzero elements. If the exclusion of the trivial case occasionally goes

unmentioned, we hope that it will be seen as an implicit assumption, and not an oversight.

It follows from Prop. 1.15 that an element Λ of $L(X,Y)$ is continuous if and only if $\|\Lambda\| < \infty$. The set of *continuous* linear mappings from X to Y is denoted by $L_C(X,Y)$. It is a vector space in its own right, and the **operator norm** $\|\cdot\|_{L(X,Y)}$ turns it into a normed space. It is easy to see that equivalent norms on X and on Y generate the same set $L_C(X,Y)$ of continuous linear mappings.

1.18 Exercise. We define a mapping $S : \ell^p \to \ell^p$ as follows: S sends $(x_1, x_2, x_3, \dots)$ to $(x_2, x_3, x_4, \dots)$. (S is a *shift to the left*.) Prove that S belongs to $L_C(\ell^p, \ell^p)$, and evaluate its operator norm. □

1.19 Proposition. *Let X be a normed space of finite dimension n, and let the set $\{e_i : i = 1, 2, \dots, n\}$ be a basis for X. Let $T : X \to \mathbb{R}^n$ be the mapping that associates to each x in X its coordinates $\lambda = (\lambda_1, \lambda_2, \dots, \lambda_n)$ with respect to this basis. Then T is continuous.*

Proof. By definition, $Tx = \lambda$ is the unique element of $\mathbb{R}^n$ for which $x = \sum_{i=1}^n \lambda_i e_i$. It follows readily that T is a linear mapping. Thus, by Prop. 1.15, it suffices to prove that $|T|$ is bounded on B_X.

We argue by contradiction. Accordingly, let x^j be a sequence of elements in B_X such that $|Tx^j|$ diverges to $+\infty$. Setting $\lambda^j = Tx^j$, we have

$$x^j = \sum_{i=1}^n \lambda_i^j e_i = \Gamma(\lambda^j), \qquad (*)$$

where Γ is a mapping of the form defined in Exer. 1.16. By taking an appropriate subsequence, we may suppose that $\lambda^j / |\lambda^j|$ converges to a unit vector $\lambda \in \mathbb{R}^n$. Dividing across by $|\lambda^j|$ in ($*$) and passing to the limit (using the continuity of Γ), we obtain

$$\lim_{j \to \infty} x^j / |\lambda^j| = 0 = \Gamma(\lambda) = \sum_{i=1}^n \lambda_i e_i .$$

This contradicts the linear independence of the e_i, and completes the proof. □

As we now proceed to show, a linear operator whose domain is finite-dimensional must be continuous.

1.20 Corollary. *If the normed space X is finite dimensional, then any $\Lambda \in L(X,Y)$ is continuous.*

Proof. Let $\{e_i : 1 \leqslant i \leqslant n\}$ be a basis for X, and let $Tx = \lambda$ be the mapping of the proposition above. Set $y_i = \Lambda e_i$. By linearity, we have

$$\Lambda x = \sum_{i=1}^n \lambda_i \Lambda e_i = \sum_{i=1}^n \lambda_i y_i = \Gamma \circ T(x) \;\; \forall\, x \in X,$$

where Γ is defined in Exer. 1.16. Then Λ, as the composition of two continuous functions, is continuous. □

Discontinuous linear functionals. It is one of the striking differences between finite and infinite dimensional normed spaces that in the latter case, there exist linear functionals which are discontinuous. A simple example of this is provided by the vector space of polynomial functions f on $[0,1]$, equipped with the norm of $C[0,1]$. If we define $\Lambda f = f'(0)$, the reader may verify that Λ is a discontinuous linear functional: the closeness of f to 0 does not imply that $f'(0)$ is close to 0.

It is possible to give an abstract construction of a discontinuous linear functional, valid in any infinite dimensional normed space X. We sketch it now. Let $\{e_i\}$ be a countable collection of linearly independent unit vectors in X. Then setting $f(e_i) = i$ for each i induces a unique linear functional on the vector subspace of X generated by $\{e_i\}$. An application of Zorn's lemma (along similar lines to the proof of Theorem 1.32) implies that f can be extended to X while remaining linear. However, f cannot be continuous at 0: the sequence e_i/i converges to 0 in X (since $\|e_i\| = 1$), yet we have $f(e_i/i) = 1 \ \forall i$.

The axiom of choice plays a role in a construction such as the above, as it does in future, rather more important ones. We may as well confess to the reader that we suffer no angst in regard to this fact.

Isometries. How can we make precise the statement that two normed spaces are essentially the same? Or that the "only" normed space of dimension n is $\mathbb{R}^n$? Such an equivalence needs to bear upon four different factors: on the underlying sets (we need a bijection), on the vector space structure (the bijection must be linear), on the topologies (it must be continuous, with continuous inverse; that is, a homeomorphism), and on the norms (which must be preserved). Thus we are led to the definition of an **isometry** between two normed spaces X and Y:

a bijective linear mapping $T : X \to Y$ such that $\|Tx\|_Y = \|x\|_X \ \forall x \in X$.

It follows from this definition that T is continuous, and that its inverse T^{-1} is continuous as well; in fact, T^{-1} is an isometry "in the other direction." We say that X and Y are *isometric*.

When two normed spaces are isometric, they are identical as normed spaces; only the labels on the points (x or Tx) change in considering one or the other.

1.21 Exercise. Show that the space $\mathrm{AC}^p[a,b]$ (see Example 1.13) is isometric to $\mathbb{R} \times L^p(a,b)$. □

The assertion that $\mathbb{R}^n$ is the only normed space of dimension n may now be given a precise meaning.

1.22 Theorem. *Let X be a normed space of finite dimension n. Then there exists an equivalent norm on X relative to which X is isometric to $\mathbb{R}^n$.*

Proof. It is understood that $\mathbb{R}^n$ is equipped with its default Euclidean norm $|\cdot|$. By hypothesis, the vector space X admits an algebraic basis consisting of n elements $\{e_1, e_2, \ldots, e_n\}$. As in Prop. 1.19, let T be the mapping such that $Tx = \lambda \in \mathbb{R}^n$ is the vector of coordinates of x with respect to the basis. Let us define $\|x\|' = |\lambda|$. Then $\|\cdot\|'$ is a norm on X, and relative to it, T is an isometry: $|Tx| = \|x\|'$.

In order to complete the proof, it suffices to show that the norm $\|\cdot\|'$ is equivalent to the original norm on X. From the relation $x = \sum_{i=1}^n \lambda_i e_i$ we deduce

$$\|x\| \leqslant \sum_{i=1}^n |\lambda_i| \|e_i\| \leqslant \big(\max_i \|e_i\|\big) n |\lambda| = c\|x\|',$$

where $c := \big(\max_i \|e_i\|\big) n$. To finish, we need to prove the existence of d such that $\|x\|' \leqslant d\|x\| \;\; \forall x$. But we know from Prop. 1.19 that T is continuous on $(X, \|\cdot\|)$, so this follows from Prop. 1.15. $\square$

There are many different topologies on $\mathbb{R}^n$, but only one norm topology:

1.23 Corollary. *Any norm on $\mathbb{R}^n$ is equivalent to the Euclidean norm.*

The following useful fact follows from Theorem 1.22.

1.24 Corollary. *A finite dimensional subspace of a normed space X is closed.*

Proof. Let S be a finite dimensional subspace of X. Then, by the theorem, there is a norm $\|\cdot\|_S$ on S such that $c\|x\|_X \leqslant \|x\|_S \leqslant d\|x\|_X \;\; \forall x \in S$, and an isometry $T : \mathbb{R}^n \to (S, \|\cdot\|_S)$. Let x_i be a sequence in S that converges in X to a limit $x \in X$. It suffices to prove that $x \in S$. To see this, note that we have

$$|T^{-1}x_i - T^{-1}x_j| = \|x_i - x_j\|_S \leqslant d\|x_i - x_j\|_X,$$

from which we deduce that $T^{-1}x_i$ is a Cauchy sequence in $\mathbb{R}^n$. Then (by a known property of $\mathbb{R}^n$), there is a point $\lambda \in \mathbb{R}^n$ such that $|T^{-1}x_i - \lambda| \to 0$. By the continuity of T, we have $\|x_i - T\lambda\|_S \to 0$. It follows that $\|x_i - T\lambda\|_X \to 0$, so that $x = T\lambda \in S$. $\square$

Compact sets. In seeking to minimize a function f over a set A, the continuity of f and the compactness of A (for some topology) are highly relevant. This is because of the well-known theorem of Weierstrass (who was a pioneer in topology, as well as in the calculus of variations) stating that a continuous function on a compact set attains its minimum.

In $\mathbb{R}^n$, compact sets abound, and are easy to characterize: they are the sets that are closed and bounded. In infinite dimensional normed spaces, however, it turns out that compact sets are rather scarce, in a manner of speaking. They do exist, of course: finite sets and closed intervals are compact, for example, as is the intersection of a closed ball with a finite dimensional subspace.

The next theorem encapsulates why infinite dimensional normed spaces are a challenge for minimization. Its proof introduces the **distance function** d_A associated to a subset A of X:

$$d_A(x) = \inf_{z \in A} \|x - z\|.$$

We remark that when A is closed, then $d_A(x) = 0$ if and only if $x \in A$.

1.25 Theorem. *The closed unit ball in a normed space X is compact if and only if X is of finite dimension.*

Proof. Let X be of finite dimension n, with norm $\|\cdot\|_1$. Then there is an equivalent norm $\|\cdot\|_2$ on X for which X is isometric to $\mathbb{R}^n$, by Theorem 1.22. But the unit ball in $\mathbb{R}^n$ is compact, and an isometry sends the ball to the ball. It follows that the ball B_2 in X relative to $\|\cdot\|_2$ is compact, as the image of a compact set by a continuous function. Then rB_2 is compact for any $r > 0$. For r sufficiently large, the ball B_1 relative to $\|\cdot\|_1$ is a closed subset of rB_2, and is therefore compact.

To prove the converse, let us suppose that X has infinite dimension. Then there exists a sequence L_n of vector subspaces of X of finite dimension (closed, by Cor. 1.24) such that $L_{n-1} \subsetneq L_n$.

Using the lemma given below, we construct a sequence x_n with $x_n \in L_n$, $\|x_n\| = 1$, and $d_{L_{n-1}}(x_n) \geqslant 1/2$. It follows that $\|x_n - x_m\| \geqslant 1/2$ for $m \neq n$. Therefore the sequence x_n admits no convergent subsequence, which proves that the unit ball is not compact.

Lemma. *Let X be a normed space and L a closed subspace, $L \neq X$. For any $\varepsilon > 0$, there exists $x \in X$ with $\|x\| = 1$ such that $d_L(x) \geqslant 1 - \varepsilon$.*

To prove the lemma, pick any $v \in X \setminus L$; we have $d := d_L(v) > 0$, since L is closed. Choose $z \in L$ such that $d \leqslant \|v - z\| \leqslant d/(1-\varepsilon)$, and set

$$x = (v - z)/\|v - z\|.$$

We claim that x is the required point. For if $y \in L$, we have

$$\|x - y\| = \left\| \frac{v - z}{\|v - z\|} - y \right\| = \frac{\|v - (z + \|v - z\| y)\|}{\|v - z\|} \geqslant \frac{d}{\|v - z\|} \geqslant 1 - \varepsilon$$

in view of the fact that $z + \|v - z\| y \in L$. □

1.26 Corollary. *A normed space contains a compact set with nonempty interior if and only if it is finite dimensional.*

1.3 The dual space

The elements of $L(X,\mathbb{R})$, as the reader has been told, are referred to as *linear functionals* (on X). The ones that happen to be continuous, that is, those lying in $L_C(X,\mathbb{R})$, constitute what is called the **dual space** of X. It is destined for great things. We denote it by X^* (often pronounced X star), and equip it with the *dual norm*: for $\zeta \in X^*$, we set

$$\|\zeta\|_* = \|\zeta\|_{X^*} = \sup\{\langle \zeta, x\rangle : x \in X, \|x\| \leqslant 1\}.$$

We concede that $\|\zeta\|_*$ is simply the usual operator norm in this special case, where the evaluation of $\zeta \in X^*$ at a point $x \in X$ has been expressed using what is called the *duality pairing* of X and X^*, denoted by $\langle \zeta, x\rangle$. The closed unit ball in X^* is denoted by $B_*(0,1)$, or just B_*.

1.27 Example. In certain cases, it is possible (and sometimes it is highly useful) to exhibit an isometry that depicts the dual space in explicit terms, as we now illustrate. Let us establish, for $1 \leqslant p < \infty$, that $(\ell^p)^*$ is isometric to ℓ^q, where $q = p_*$ is the conjugate exponent to p.

We begin by observing that any sequence $v = (v_1, v_2, \dots)$ in ℓ^q induces an element ζ_v in $(\ell^p)^*$ by means of the definition

$$\langle \zeta_v, u\rangle = \sum_{i=1}^{\infty} v_i u_i, \quad u \in \ell^p.$$

This definition makes sense (that is, the sum is well defined) as a result of Hölder's inequality, which yields $|\langle \zeta_v, u\rangle| \leqslant \|v\|_q \|u\|_p$. It follows that the dual norm of ζ_v is no greater than $\|v\|_q$. When we take u to be the element of ℓ^p whose i-th term is $v_i|v_i|^{q-2}$ (or 0 if $v_i = 0$), then we obtain equality in the preceding:

$$|\langle \zeta_v, u\rangle| = \sum_{i=1}^{\infty} |v_i|^q = \|v\|_q^q = \|v\|_q \|u\|_p,$$

whence $\|\zeta_v\|_* = \|v\|_q$.

It follows that the linear application T which maps $v \in \ell^q$ to $\zeta_v \in (\ell^p)^*$ preserves the norm, which implies that it is continuous and injective. To see that it constitutes the sought-for isometry, we need only check that it is onto. We proceed to do this for the case $p \neq 1$ (so that q is finite).

Let e_i be the element of ℓ^p whose terms are all zero except for the i-th, which equals 1. Let ζ be any element of $(\ell^p)^*$, and define w to be the sequence whose i-th term is $\langle \zeta, e_i\rangle$. We claim that w belongs to ℓ^q. To see this, let u^n be the element of ℓ^p whose i-th term is $w_i|w_i|^{q-2}$ for $i \leqslant n$, and 0 for $i > n$. Applying ζ to u^n, we deduce

$$\langle \zeta, u^n \rangle = \Big\langle \zeta, \sum_1^n u_i^n e_i \Big\rangle = \sum_1^n |w_i|^q \leqslant \|\zeta\|_* \|u^n\|_p$$
$$= \|\zeta\|_* \Big\{ \sum_1^n |w_i|^{(q-1)p} \Big\}^{1/p} = \|\zeta\|_* \Big\{ \sum_1^n |w_i|^q \Big\}^{1/p}.$$

We derive from this the inequality

$$\Big\{ \sum_1^n |w_i|^q \Big\}^{1-1/p} = \Big\{ \sum_1^n |w_i|^q \Big\}^{1/q} \leqslant \|\zeta\|_*.$$

Since the right side is independent of n, it follows that w belongs to ℓ^q, as claimed.

We have $\langle \zeta, u \rangle = \langle Tw, u \rangle$ for all $u \in \ell_c^\infty$, a set which is evidently dense in ℓ^p. By the continuity of ζ and Tw, we deduce $\zeta = Tw$, proving that T is onto. □

1.28 Exercise. Complete the argument above in the case $p = 1$. □

Dual spaces of Cartesian products are easily characterized; in a certain sense, we can identify $\big(X \times Y\big)^*$ with $X^* \times Y^*$.

1.29 Proposition. *Let $Z = X \times Y$ be the Cartesian product of two normed spaces, where Z is equipped with one of the usual product norms. Then $\zeta \in Z^*$ if and only if there exist $\zeta_1 \in X^*$ and $\zeta_2 \in Y^*$ such that*

$$\langle \zeta, (x,y) \rangle = \langle \zeta_1, x \rangle + \langle \zeta_2, y \rangle, \quad x \in X, y \in Y.$$

The proof is omitted, as is the proof of the next result.

1.30 Proposition. *The dual of $\mathbb{R}^n$ is isometric to $\mathbb{R}^n$: every element ζ of $(\mathbb{R}^n)^*$ admits a unique $u \in \mathbb{R}^n$ such that $\langle \zeta, x \rangle = u \bullet x \ \ \forall x \in \mathbb{R}^n$. Then $\|\zeta\|_* = |u|$, and the map $\zeta \to u$ is an isometry.*

Notation. The notation $u \bullet x$ above refers of course to the usual dot product in $\mathbb{R}^n$. Since the dual of $\mathbb{R}^n$ can be identified with $\mathbb{R}^n$ itself, we also write $\langle u, x \rangle$ (the effect of u on x, or vice versa) for the dot product.

When the dual of a space is isometric to the space itself, as above, then to know the dual is to know the space. But this is not the case in general, and the question of whether the dual determines the space is a subtle one that will be taken up later.

1.31 Exercise. Let $1 \leqslant p \leqslant \infty$, and let $u \in L^p(\Omega)$. Let q be the conjugate exponent to p. Show that

$$\langle T_u, g \rangle := \int_\Omega u(x)\, g(x)\, dx, \ \ g \in L^q(\Omega)$$

defines an element T_u in the dual of $L^q(\Omega)$, and that $\|T_u\|_{L^q(\Omega)^*} = \|u\|_{L^p(\Omega)}$. □

The characterization of the dual space of a subspace L of X is more delicate than that of a product; describing the relationship between L^* and X^* depends upon knowing that continuous linear functionals on L can be extended continuously to X.

The following famous theorem was designed for precisely that purpose. Note that the norm on X plays no role here for the moment.

1.32 Theorem. (Hahn-Banach extension) *Let X be a vector space, and $p : X \to \mathbb{R}$ a function satisfying*

(a) $p(tx) = t\,p(x)\ \ \forall x \in X, t \geqslant 0$ (positive homogeneity);

(b) $p(x+y) \leqslant p(x)+p(y)\ \ \forall x, y \in X$ (subadditivity).

Let L be a linear subspace of X, and let $\lambda : L \to \mathbb{R}$ be a linear functional such that $\lambda \leqslant p$ on L. Then there exists a linear functional Λ on X which extends λ (that is, $\Lambda(x) = \lambda(x)\ \ \forall x \in L$) and which satisfies $\Lambda \leqslant p$ on X.

Proof. The proof is based upon an application of Zorn's lemma.[3] We consider the set P of all couples $(D(h), h)$ where $D(h)$ is a subspace of X which contains L, and where $h : D(h) \to \mathbb{R}$ is a linear functional which extends λ and which satisfies $h \leqslant p$ on $D(h)$. Note that P is nonempty, since it contains (L, λ).

We write $(D(h_1), h_1) \leqslant (D(h_2), h_2)$ when $D(h_1) \subset D(h_2)$ and h_2 extends h_1. This defines a partial order on P. If $\{(D(h_\alpha), h_\alpha)\}_\alpha$ is a totally ordered subset of P, then a majorant $(D(h), h)$ for the set is obtained by setting

$$D(h) = \bigcup_\alpha D(h_\alpha), \quad h(x) = h_\alpha(x) \text{ when } x \in D(h_\alpha).$$

The reader may verify that $D(h)$ is a subspace and that h is well defined, as a consequence of the fact that the subset is totally ordered.

Thus the set P is inductive, and by Zorn's lemma, it admits a maximal element $(D(\Lambda), \Lambda)$. If $D(\Lambda) = X$, then Λ is the extension we seek. To conclude the proof, we suppose that there is a point $x_0 \in X \setminus D(\Lambda)$, and we proceed to derive a contradiction.

Let $D(h)$ be the vector subspace of X generated by $D(\Lambda) \cup \{x_0\}$; every element of $D(h)$ is expressible (uniquely) in the form $x+tx_0$, where $x \in D(\Lambda)$ and $t \in \mathbb{R}$. We define $h : D(h) \to \mathbb{R}$ by

$$h(x+tx_0) = \Lambda(x)+\beta t,$$

where the scalar β remains to be determined. Whatever choice of β is made, the mapping h extends Λ. But in order to have $h \leqslant p$ on $D(h)$ (so that $(D(h), h)$ is an

---

[3] Let Q be a subset of a partially ordered set $(P, \leqslant)$. A *majorant* of Q is an element $p \in P$ such that $q \leqslant p\ \ \forall q \in Q$. The set Q is *totally ordered* if every pair x, y of points in Q satisfies either $x \leqslant y$ or $y \leqslant x$. P is *inductive* if every totally ordered subset Q of P admits a majorant. Zorn's lemma affirms that every (nonempty) inductive partially ordered set P admits a *maximal* element: a point m such that $x \in P,\ m \leqslant x$ implies $x = m$.

element of P), we must have

$$\Lambda(x)+\beta t \leqslant p(x+tx_0) \;\; \forall t \in \mathbb{R},\; x \in D(\Lambda).$$

Let $t > 0$. Then the inequality above can be divided by t, and (using the linearity of Λ and the positive homogeneity of p) is seen to be equivalent to

$$\Lambda(x)+\beta \leqslant p(x+x_0) \;\; \forall x \in D(\Lambda),$$

since either x or x/t can be used to indicate a generic element of $D(\Lambda)$. In other words, the required inequality for $t > 0$ reduces to the case $t = 1$. The case $t < 0$ is treated similarly, by reduction to $t = -1$. We summarize: to have $h \leqslant p$ on $D(h)$, it is necessary and sufficient that

$$\Lambda(x)+\beta \leqslant p(x+x_0) \;\; \forall x \in D(\Lambda), \quad \Lambda(y)-\beta \leqslant p(y-x_0) \;\; \forall y \in D(\Lambda).$$

To put this yet another way, we require that β satisfy

$$\sup_{y \in D(\Lambda)} \Lambda(y)-p(y-x_0) \leqslant \beta \leqslant \inf_{x \in D(\Lambda)} p(x+x_0)-\Lambda(x).$$

Such a choice of β is possible if and only if

$$\Lambda(x)+\Lambda(y) \leqslant p(x+x_0)+p(y-x_0) \;\; \forall x,y \in D(\Lambda).$$

But the left side of this inequality coincides with $\Lambda(x+y)$,which is bounded above by $p(x+y)$, which, in turn, is bounded above by the right side (in view of the subadditivity of p). Thus $(D(h),h)$ belongs to P (for a suitable choice of β), contradicting the maximality of $(D(\Lambda),\Lambda)$. □

1.33 Corollary. *Let L be a linear subspace of the normed space X, and let $\lambda : L \to \mathbb{R}$ be a continuous linear functional. Then there exists $\Lambda \in X^*$ extending λ such that $\|\Lambda\|_{X^*} = \|\lambda\|_{L^*}$.*

Proof. Use the theorem with $p(x) = \|\lambda\|_{L^*}\|x\|$. □

The corollary above spawns one of its own, which resolves the problem of describing the dual space of a subspace:

1.34 Corollary. *Let L be a subspace of the normed space X. Then the dual space of L consists of the restrictions to L of elements of the dual of X:*

$$L^* = \left\{ \zeta|_L : \zeta \in X^* \right\}.$$

1.35 Exercise. Given any $x_0 \in X$, prove the existence of $\zeta_0 \in X^*$ such that

$$\|\zeta_0\|_* = \|x_0\|, \quad \langle \zeta_0, x_0 \rangle = \|x_0\|^2.$$

Deduce the formula $\|x\| = \max\left\{ \langle \zeta, x \rangle : \|\zeta\|_* \leqslant 1 \right\} \;\; \forall x \in X$. □

1.4 Derivatives, tangents, and normals

It is a very useful technique to approximate the nonlinear by the linear. For functions, this involves the use of derivatives. For sets, it is tangent and normal vectors which are involved. We proceed to introduce these basic constructs, beginning with the one the reader knows best.

Derivatives. Let $F : X \to Y$ be a mapping between two normed spaces. The derivative of F at a point x, when it exists, is an element $F'(x)$ of $L_C(X,Y)$ such that

$$\lim_{\substack{u \to x \\ x \neq u}} \frac{\| F(u) - F(x) - \langle F'(x), u - x \rangle \|_Y}{\| u - x \|_X} = 0.$$

We say that F is *differentiable* at x when such a mapping $F'(x)$ exists; in that case, it is unique. Differentiability[4] does not depend on the choice among equivalent norms. A function which is differentiable at x is necessarily continuous at x.

Notation: The derivative $F'(x)$ is also denoted $DF(x)$ at times. If F is a function of two variables x and u, the derivative of the function $F(\cdot, u)$ at x (for a given value of u) is denoted by either $D_x F(x,u)$ or $F'_x(x,u)$.

We summarize below some familiar facts from differential calculus; the usual proofs adapt without difficulty to the setting of normed spaces.

1.36 Proposition. *Let X, Y, and Z be normed spaces.*

(a) (Linearity) *Let F and H be functions from X to Y, each differentiable at x, and let $c, k \in \mathbb{R}$. Then $cF + kH$ is differentiable at x, and*

$$(cF + kH)'(x) = cF'(x) + kH'(x).$$

(b) (Chain rule) *Let $F : X \to Y$ be differentiable at x and let $\theta : Y \to Z$ be a function which is differentiable at $F(x)$. Then $\theta \circ F$ is differentiable at x, and we have*

$$(\theta \circ F)'(x) = \theta'\big(F(x)\big) \circ F'(x).$$

(c) (Fermat's rule) *Let $f : X \to \mathbb{R}$ attain a local minimum at x, a point at which f is differentiable. Then $f'(x) = 0$.*

(d) (Mean value theorem) *Let $f : X \to \mathbb{R}$ be continuous at each point of the segment $[u,v]$, and differentiable at each point of (u,v). Then there exists $w \in (u,v)$ such that*

$$f(v) - f(u) = \langle f'(w), v - u \rangle.$$

[4] Strictly speaking, $F'(x)$ is known as the *Fréchet derivative*, to distinguish it from (surprisingly many!) other objects of similar type.

The function $F : X \to Y$ is said to be *continuously differentiable* at x if $F'(u)$ exists for all u in a neighborhood of x, and if the mapping $u \mapsto F'(u)$ is continuous at x. The continuity referred to here is that of a mapping between the normed spaces X and $L_C(X,Y)$; it is equivalent to requiring

$$\lim_{u \to x} \left\| F'(u) - F'(x) \right\|_{L_C(X,Y)} = 0.$$

Given a function $F : X \to Y$ and $v \in X$, the **directional derivative** of F at x in the direction v, denoted $F'(x;v)$, refers to the following limit (when it exists):

$$F'(x;v) = \lim_{t \downarrow 0} \frac{F(x+tv) - F(x)}{t}.$$

Functions can admit directional derivatives even when they fail to be differentiable. (Consider $f(x) = |x|$ at $x = 0$.) Thus, the introduction of directional derivatives is a step toward requiring less smoothness of the data. It is easy to show that if F is differentiable at x, then F has directional derivatives at x given by

$$F'(x;v) = \langle F'(x), v \rangle, \quad v \in X.$$

Thus, in this case, the mapping $v \mapsto F'(x;v)$ turns out to be linear; in general, it is merely positively homogeneous:

$$F'(x;tv) = t\,F'(x;v), \quad t > 0.$$

The tangent and normal cones. Roughly speaking, it may be said that tangent vectors to a set, at a point x in the set, correspond to "directions that stay in the set" as one follows them from x. Normal vectors, on the other hand, correspond to "orthogonal" directions which *leave* the set in an efficient way.

In the classical setting of smooth manifolds, tangent vectors form a linear subspace; this is analogous to the derivative being a linear mapping. In contexts that are less smooth, then, just as we replace the derivative by a (positively homogeneous) directional derivative, so it is that the tangent space is replaced by the more general construct known as a **cone**. A cone K in a vector space is a nonempty set which is stable under multiplication by a positive scalar: $v \in K,\ t > 0 \implies tv \in K$. Thus, a subspace is a cone, but the converse fails; in $\mathbb{R}^2$, for example, the first quadrant is a cone. Note that a closed cone necessarily contains 0.

Let S be a subset of X. The **tangent cone** to S at a point $x \in S$, denoted $T_S(x)$, consists of all points $v \in X$ expressible in the form

$$v = \lim_{i \to \infty} \frac{x_i - x}{t_i},$$

where x_i is a sequence in S converging to x, and t_i is a positive sequence decreasing to 0. The reader should note that, like the derivative of a function, the tangent cone $T_S(x)$ is a *local* construct: it depends only upon the nature of the set S in a neigh-

borhood of the point x. We remark that $T_S(x)$ is also referred to as the Bouligand tangent cone.

1.37 Exercise. Let x be a point in S.

(a) Prove that $T_S(x)$ is a closed cone.

(b) Show that an alternate definition of tangency is provided by the following:

$v \in T_S(x)$ if and only if there is a sequence v_i in X converging to v, and a positive sequence t_i decreasing to 0, such that $x + t_i v_i \in S \ \forall i$.

(c) Prove that $v \in T_S(x)$ if and only if $\liminf_{t \downarrow 0} d_S(x+tv)/t = 0$. (Recall that d_S refers to the distance function of S.) □

The **normal cone** $N_S(x)$ to S at x is the subset of the dual space defined by

$$N_S(x) = \big\{ \zeta \in X^* : \langle \zeta, v \rangle \leqslant 0 \ \ \forall v \in T_S(x) \big\}.$$

We say that the normal cone is obtained from the tangent cone by *polarity*. The polar of any subset A of X is defined and denoted as follows:

$$A^{\triangle} = \big\{ \zeta \in X^* : \langle \zeta, x \rangle \leqslant 0 \ \ \forall x \in A \big\}.$$

(Polarity is studied in detail in §4.3.) It is apparent from the definition that $N_S(x)$ is a closed cone in X^*. Note also that

$$x \in \operatorname{int} S \implies T_S(x) = X, \quad N_S(x) = \{0\}.$$

1.38 Exercise.

(a) Let K be a closed cone in X. Prove that $T_K(0) = K$.

(b) Let L be a closed subspace of X. Show that, for any $x \in L$, we have

$$T_L(x) = L \ \text{ and } \ N_L(x) = \big\{ \zeta \in X^* : \langle \zeta, x \rangle = 0 \ \ \forall x \in L \big\}.$$

(c) Let A and E be subsets of normed spaces X and Y respectively, and let (x,y) belong to $A \times E$. Prove that

$$T_{A\times E}(x,y) = T_A(x) \times T_E(y), \quad N_{A\times E}(x,y) = N_A(x) \times N_E(y).$$ □

Optimization. The next result shows how the normal cone arises in expressing Fermat's rule in a context of constrained minimization.

1.39 Proposition. *Let $f : X \to \mathbb{R}$ be differentiable, and let f attain a minimum over the set A at a point x. Then we have*

$$-f'(x) \in N_A(x) \ \text{ and } \ \langle f'(x), v \rangle \geqslant 0 \ \ \forall v \in T_A(x).$$

Proof. We proceed to prove the second assertion. Let v belong to $T_A(x)$. Then, according to Exer. 1.37, there exists a sequence v_i in X converging to v, and a positive sequence t_i decreasing to 0, such that, for each index i, we have $x + t_i v_i \in A$. The optimality of x implies

$$f(x + t_i v_i) - f(x) \geqslant 0 \;\; \forall\, i\,.$$

Dividing this inequality across by t_i and passing to the limit as $i \to \infty$, we obtain $\langle f'(x), v \rangle \geqslant 0$, which confirms the second assertion of the proposition.

The first assertion follows immediately from the second, and in fact is equivalent to it, in view of the way $N_A(x)$ is defined through polarity with $T_A(x)$. □

Derivatives, tangents, and normals come together when we consider sets that are defined by functional relations, as in the following.

1.40 Exercise. Let X and Y be normed spaces, and let S be given by

$$S = \big\{ u \in X : F(u) = 0 \big\},$$

where the function $F : X \to Y$ is differentiable at a point $x \in S$. Prove that

$$T_S(x) \subset \big\{ v \in X : \langle F'(x), v \rangle = 0 \big\}, \;\; N_S(x) \supset \big\{ \zeta = \Lambda \circ F'(x) : \Lambda \in Y^* \big\}. \;\; □$$

The following construct is frequently used in formulas such as the last one above.

Adjoint operators. Let X and Y be normed spaces and $T \in L_C(X,Y)$. It is not difficult to see that one defines a unique element T^* of $L_C(Y^*, X^*)$ by the formula

$$\langle T^* y^*, x \rangle = \langle y^*, Tx \rangle, \quad x \in X,\; y^* \in Y^*.$$

T^* is called the *adjoint* of the operator T. One often meets the adjoint in differential calculus, notably in connection with the chain rule. In Prop. 1.36, the latter was expressed in the form

$$(\theta \circ F)'(x) = \theta'\big(F(x)\big) \circ F'(x).$$

Using the adjoint, one may write instead

$$(\theta \circ F)'(x) = \langle F'(x)^*, \theta'\big(F(x)\big) \rangle.$$

In terms of the adjoint, the second estimate in Exer. 1.40 asserts $N_S(x) \supset F'(x)^* Y^*$.

The issue of when equality holds in such estimates is a delicate one related to implicit functions; it will be studied later, together with more general cases in which $T_S(x)$ is a cone rather than a subspace. We shall see as well that normal and tangent cones enjoy a calculus of their own, as hinted at by the following.

1.41 Exercise. Let $x \in S_1 \cap S_2$, where S_1 and S_2 are closed subsets of X. Prove that

$$T_{S_1 \cap S_2}(x) \subset T_{S_1}(x) \cap T_{S_2}(x), \quad N_{S_1 \cap S_2}(x) \supset N_{S_1}(x) + N_{S_2}(x). \qquad □$$

Tangents and normals in Euclidean space. In $\mathbb{R}^n$, the normal and tangent cones can be thought of as living in the same space (since we identify the dual of $\mathbb{R}^n$ with itself). This allows us to picture the geometry of tangent and normal cones more intuitively.

1.42 Exercise.

(a) Find the normal cone to the interval $[-1,1] \subset \mathbb{R}$ at the point -1, and at 1.

(b) Find the normal cone to the cube $[-1,1]^3 \subset \mathbb{R}^3$ at the point $(-1,0,1)$. □

The following confirms our understanding of the normal cone as consisting of directions that leave the set.

1.43 Proposition. *Let $x \in S \subset \mathbb{R}^n$. Then $T_S(x) \cap N_S(x) = \{0\}$. For any nonzero $\zeta \in N_S(x)$, for all $r > 0$ sufficiently small, we have $x + r\zeta \notin S$. More generally, there exists $r > 0$ such that*

$$S \cap \big(x + N_S(x)\big) \cap B(x,r) = \{x\}.$$

Proof. Let v belong to both $T_S(x)$ and $N_S(x)$. Then, by definition of the normal cone, we have $v \bullet v \leqslant 0$, whence $v = 0$, which proves the first assertion. We now prove the final assertion, which can easily be seen to subsume the other. We argue by contradiction.

If the desired conclusion fails, then there is a sequence of nonzero elements ζ_i in $N_S(x)$ converging to 0 such that $x + \zeta_i \in S \ \ \forall i$. By taking a subsequence, we may suppose that $\zeta_i / |\zeta_i|$ converges to a unit vector ζ. Since $N_S(x)$ is a closed cone, we have $\zeta \in N_S(x)$. It follows from the definition of tangent cone that

$$\zeta = \lim_{i \to \infty} (x + \zeta_i - x)/|\zeta_i| \in T_S(x).$$

Then $\zeta \in T_S(x) \cap N_S(x)$, whence $\zeta = 0$, which is the required contradiction. □

Manifolds. The sets S we meet in practice are sometimes defined as *level sets*. This means that S consists of the points $u \in \mathbb{R}^n$ which satisfy the equation $F(u) = 0$, for some given function $F : \mathbb{R}^n \to \mathbb{R}^k$. The reader will have encountered the classical case of such a set in which F is continuously differentiable, $1 \leqslant k < n$, and the following *rank condition* is postulated: the $k \times n$ Jacobian matrix $DF(x)$ has maximal rank (that is, rank k) at each point x in S. Then S is called a *manifold*.[5] We shall establish later (§5.4) that for manifolds, the two estimates appearing in Exer. 1.40 hold with equality. In that case, the cones $T_S(x)$ and $N_S(x)$ turn out to be orthogonal subspaces of $\mathbb{R}^n$ of dimension $n - k$ and k respectively. The former is the null space of the matrix $DF(x)$, while the latter is its row space.

[5] We have used the same notation $DF(x)$ for the Jacobian as for the derivative of F. This is justified by the fact that the linear mapping from $\mathbb{R}^n$ to $\mathbb{R}^k$ which the matrix induces by matrix multiplication (on the left) is precisely the derivative of F at x.

1.44 Example. The following modest portfolio of examples illustrates the nature of tangents and normals in various cases.

The first set in Fig. 1.1 below corresponds to a manifold defined as above, where

$$n = 2,\ k = 1,\ F(x,y) = x^2 + y^2 - 1.$$

Thus, S is the unit circle here. As indicated, the tangent space at the point $(1,0)$ is the y-axis (so to speak), and the normal space is its orthogonal complement, namely the x-axis.

The second set in Fig. 1.1 is also smooth, but it is defined now by an *inequality* $x^2 + y^2 - 1 \leqslant 0$; that is, S is the unit disc. This is an instance of a *manifold with boundary*, and its tangent cone at $(1,0)$ is the halfspace $\{\,(x,y) : x \leqslant 0\,\}$. The normal cone is the single ray consisting of the positive x-axis.

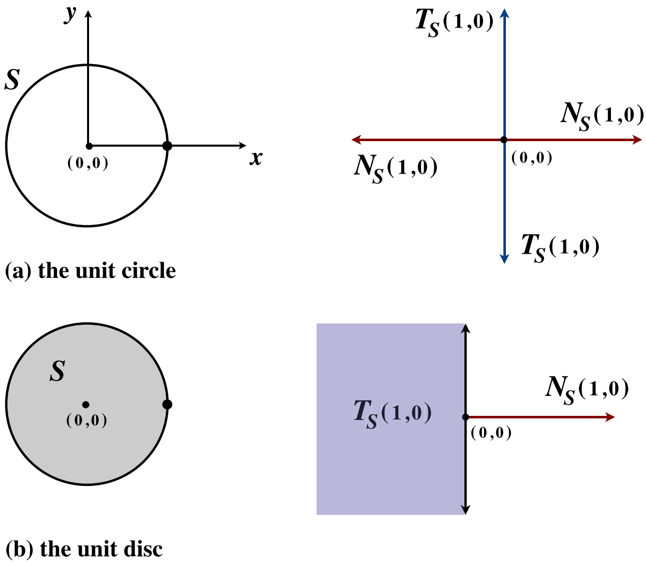
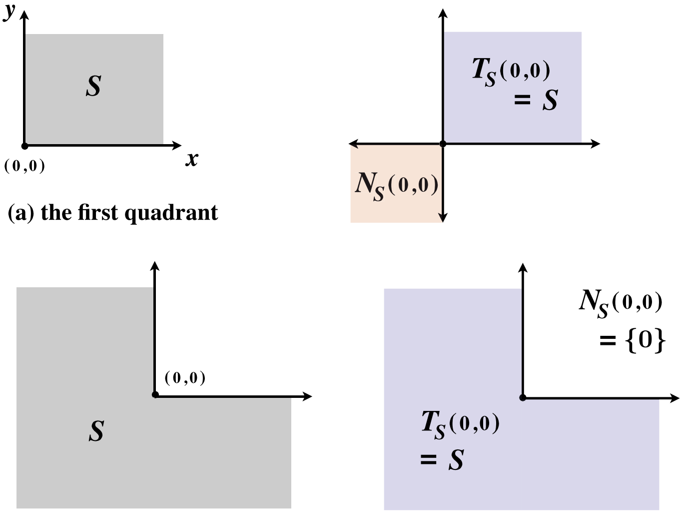


Fig. 1.1
The unit circle, the unit disc, their tangent and normal cones at $(1,0)$.

In Fig. 1.2, the reader finds two examples of *nonsmooth* sets. In (a), S is the first quadrant. This is already a closed cone, so it coincides with its tangent cone at $(0,0)$. Applying polarity, we find that the normal cone at $(0,0)$ is given by the third quadrant.

The first set S in Fig. 1.2 is *convex*, an important class of sets whose acquaintance the reader is just about to make. We shall see later that a direct definition of normal cone (without reference to tangents) can be given for convex sets.



(b) the closure of its complement

Fig. 1.2
The first quadrant, the closure of its complement, their tangent and normal cones at $(0,0)$.

The set in Fig. 1.2 (b), on the other hand, is neither smooth nor convex (it is the closure of the complement of the first quadrant). Once again, its tangent cone at the origin is the set itself. The normal cone, however, reduces here to $\{0\}$; we say then that it is *trivial*. We know that the normal cone always contains 0, from the way it is defined. □

Chapter 2
Convex sets and functions

The class of convex sets plays a central role in functional analysis. The reader may already know that a subset C of the vector space X is said to be **convex** if the following implication holds:

$$t \in (0,1), \; x, y \in C \implies (1-t)x + ty \in C.$$

Thus, a convex set is one that always contains the segment between any two of its points. When X is a normed space, as we assume in this chapter, the triangle inequality implies that open or closed balls are convex; the complement of a ball fails to be convex. A moment's thought confirms as well that convex sets share an important property of closed sets: *an arbitrary intersection of convex sets is convex.*

As we shall see, the possibility of *separating* two disjoint convex sets by a hyperplane is a fundamental issue. The celebrated Hahn-Banach theorem, that bears on this point, is perhaps the single most useful tool in the classical theory. We shall also meet convex *functions* in this chapter. These counterparts of convex sets turn out to be equally important to us later on.

2.1 Properties of convex sets

A **convex combination** of finitely many points $x_1, x_2, \ldots, x_m$ in X ($m \geqslant 2$) means any element x of the form

$$x = \sum_{i=1}^{m} t_i x_i,$$

where the *coefficients* t_i of the convex combination are nonnegative numbers summing to one: $t_i \geqslant 0$ and $\sum_{i=1}^{m} t_i = 1$.

2.1 Exercise. The set C is convex if and only if any convex combination of points in C belongs to C. □

F. Clarke, *Functional Analysis, Calculus of Variations and Optimal Control*,
Graduate Texts in Mathematics 264, DOI 10.1007/978-1-4471-4820-3_2,

The following facts are essential, not surprising, yet tricky to prove if one doesn't go about it just right.

2.2 Theorem. *Let C be a convex subset of the normed space X. Then*

(a) $\overline{C}$ *is convex*;

(b) $x \in \overline{C},\ y \in C^{\circ} \implies (x,y] \subset C^{\circ}$;

(c) C° *is convex*;

(d) $C^{\circ} \neq \emptyset \implies \overline{C} = \overline{C^{\circ}}$ *and* $C^{\circ} = \big(\overline{C}\big)^{\circ}$.

Proof. Let $x, y \in \overline{C}$ and $0 < t < 1$. In order to prove that the point $(1-t)x+ty$ lies in $\overline{C}$ (thus proving (a)), we must show that, given any neighborhood U of 0, the set $(1-t)x+ty+U$ meets C (that is, has nonempty intersection with C). Let V be a neighborhood of 0 such that $(1-t)V+tV \subset U$. Then we have

$$(1-t)x+ty+U \supset (1-t)(x+V)+t(y+V).$$

But $x+V$ and $y+V$ both meet C, since x and y belong to $\overline{C}$. Because C is convex, we obtain the desired conclusion.

We turn now to (b). Let $0 < t < 1$; we wish to show that $(1-t)x+ty \in C^{\circ}$. There is a neighborhood V of 0 satisfying $y+V+V \subset C$. Furthermore, $x-tV/(1-t)$ meets C, so there exist $v \in V$, $c \in C$ such that $x = c+tv/(1-t)$. We then find

$$(1-t)x+ty+tV = (1-t)c+t(y+v+V) \subset (1-t)c+tC \subset C,$$

in light of the convexity of C. Since tV is a neighborhood of 0, the conclusion follows.

The reader will observe that part (c) of the theorem follows immediately from (b). We turn then to (d). Let y be a point in C°, and $x \in \overline{C}$. By part (b), we have the containment $(x,y] \subset C^{\circ}$; this implies $x \in \overline{C^{\circ}}$. The inclusion $\overline{C} \supset \overline{C^{\circ}}$ being evident, the first assertion follows. To prove the other, it suffices to show that

$$x \in \big(\overline{C}\big)^{\circ} \implies x \in C^{\circ}.$$

If x belongs to the set on the left, there is a neighborhood V of 0 such that $x+V \subset \overline{C}$, whence, by part (b), we have

$$(1-t)(x+V)+ty \subset C^{\circ} \ \forall t \in (0,1].$$

Now for $t > 0$ sufficiently small, the point $t(x-y)/(1-t)$ belongs to V. For such a value of t, we deduce

$$x = (1-t)\Big(x+\frac{t(x-y)}{1-t}\Big)+ty \in (1-t)(x+V)+ty \subset C^{\circ}.$$

□

2.3 Exercise. Show that the hypothesis $C^{\circ} \neq \emptyset$ in the last assertion of Theorem 2.2 is required for the conclusion, as well as the convexity of C. □

Convex envelopes. Let S be a subset of X. The *convex envelope* of S, denoted $\operatorname{co} S$, is the smallest convex subset of X containing S. This definition is meaningful, since there is at least one convex set containing S (the space X itself), and since the intersection of convex sets is convex; thus, $\operatorname{co} S$ is the intersection of all convex sets containing S. The convex envelope of S can also be described as the set of all convex combinations generated by S:

2.4 Exercise. Show that

$$\operatorname{co} S = \left\{ \sum_{i=1}^{m} t_i x_i : m \geqslant 1,\, x_i \in S,\; t_i \geqslant 0,\; \sum_{i=1}^{m} t_i = 1 \right\},$$

and deduce that $\operatorname{co}(S_1 + S_2) \subset \operatorname{co} S_1 + \operatorname{co} S_2$. □

The *closed convex envelope* of S is the smallest closed convex set containing S. It is denoted by $\overline{\operatorname{co}}\, S$. Clearly, it corresponds to the intersection of all closed convex sets containing S.

2.5 Exercise. Prove that $\overline{\operatorname{co}}\, S = \operatorname{cl}(\operatorname{co} S)$. □

The characterization of $\operatorname{co} S$ given in Exer. 2.4 involves arbitrarily large integers m. In finite dimensions, however, this can be improved upon:

2.6 Proposition. (Carathéodory's theorem) *Let S be a subset of a normed space X of finite dimension n. Let $x \in \operatorname{co} S$. Then there is a subset A of S containing at most $n+1$ points such that x is a convex combination of the points of A.*

Proof. Let $x = \sum_0^k t_i x_i$ be a convex combination of $k+1$ elements of S for $k > n$. We proceed to show that x is, in fact, the convex combination of k of these elements, which implies the result.

We may suppose $t_i > 0$, $0 \leqslant i \leqslant k$, for otherwise there is nothing to prove. The vectors $x_i - x_0$ $(1 \leqslant i \leqslant k)$ are linearly dependent in X, since $k > n$. There exist, therefore, scalars r_i $(1 \leqslant i \leqslant k)$, not all zero, such that $\sum_1^k r_i(x_i - x_0) = 0$. Now define $r_0 = -\sum_1^k r_i$. Then we have $\sum_0^k r_i = 0$, $\sum_0^k r_i x_i = 0$. We pick an index j for which r_i/t_i is maximized:

$$r_i/t_i \leqslant r_j/t_j\,, \quad i = 0, 1, \ldots, k.$$

Then $r_j > 0$ (since the r_i are not all zero and sum to zero). We proceed to set

$$c_i = t_i - r_i t_j / r_j\,, \quad 0 \leqslant i \leqslant k.$$

We then find $c_i \geqslant 0$, $\sum_0^k c_i = 1$, $x = \sum_0^k c_i x_i$ as well as $c_j = 0$, which expresses x in the required way. □

2.7 Exercise. In $\mathbb{R}^2$, let S consist of the points (x,y) on the unit circle that lie in the first quadrant, together with the points $(-1,0)$ and $(0,-1)$. Certain points in $\operatorname{co} S$ can be expressed as a convex combination of two points in S; others require three. Which points require three? □

2.8 Exercise. Let S be a compact subset of $\mathbb{R}^n$. Prove that $\operatorname{co} S$ is compact. □

When the underlying set is convex, the tangents and normals that we met in §1.4 admit alternate characterizations, as we now see. The reader may wish to ponder these in connection with the two of the four sets in Figures 1.1 and 1.2 (pp. 24–25) that are convex.

2.9 Proposition. *Let S be a convex set in X, and let $x \in S$. Then $T_S(x)$ is convex, $S \subset x + T_S(x)$, and we have*

$$T_S(x) = \operatorname{cl}\Big\{ \frac{u-x}{t} : t > 0,\, u \in S \Big\},\; N_S(x) = \big\{ \zeta \in X^* : \langle \zeta, u-x \rangle \leqslant 0 \;\; \forall u \in S \big\}.$$

Proof. First, we call upon the reader to show (with the help of Theorem 2.2) that the following set W is convex:

$$\operatorname{cl}\Big\{ \frac{u-x}{t} : t > 0,\; u \in S \Big\}.$$

It is clear from the definition of tangent vector that $T_S(x) \subset W$. To prove the opposite inclusion, it suffices to show that any vector of the form $v = (u-x)/t$, where u is in S and $t > 0$, belongs to $T_S(x)$, since the latter is closed. We do this now.

Let ε_i be a positive sequence decreasing to 0. Then, for i sufficiently large, the point $x_i = x + \varepsilon_i(u-x)$ belongs to S, since S is convex. For such i, we have v equal to $(x_i - x)/(t\varepsilon_i)$. This (constant) sequence converges to v, which makes it clear that $v \in T_S(x)$ by definition of the tangent cone. The characterization of $T_S(x)$ is therefore proved; clearly, it implies $S \subset x + T_S(x)$. Finally, since the normal cone is defined by polarity with respect to the tangent cone, the stated expression for $N_S(x)$ is a direct consequence of that characterization as well. □

2.2 Extended-valued functions, semicontinuity

It will be very useful for later purposes to consider functions with values in the **extended reals**; that is, functions $f : X \to \mathbb{R} \cup \{+\infty\}$. The reader is not to suppose by this that we are slipping into informality; the idea is to accommodate such extended-valued functions without lowering our customary standards of rigor.

Notation and terminology: we denote $\mathbb{R} \cup \{+\infty\}$ by $\mathbb{R}_\infty$. The **effective domain** of an extended-valued function f, denoted $\operatorname{dom} f$, is the set

$$\operatorname{dom} f = \{x \in X : f(x) < \infty\}.$$

The function f is called **proper** when $\operatorname{dom} f \neq \emptyset$. The **epigraph** of f is the set of points on or above the graph:

$$\operatorname{epi} f = \{(x,r) \in X \times \mathbb{R} : f(x) \leqslant r\}.$$

The following two types of extended-valued functions play an important role.

2.10 Definition. *Let Σ be a nonempty subset of X^*. The* **support function** *of Σ is the mapping $H_\Sigma : X \to \mathbb{R}_\infty$ defined by*

$$H_\Sigma(x) = \sup_{\sigma \in \Sigma} \langle \sigma, x \rangle, \ x \in X.$$

Let S be a subset of X. Its **indicator function** *$I_S : X \to \mathbb{R}_\infty$ is the function which has value 0 on S and $+\infty$ elsewhere.*

It is unreasonable to ask of such functions that they be continuous. A more appropriate regularity property for these and other extended-valued functions that we shall encounter, is the following. We state it in an arbitrary topological space.

2.11 Definition. *Let E be a set endowed with a topology. A function $f : E \to \mathbb{R}_\infty$ is said to be* **lower semicontinuous** *(abbreviated* lsc*) if, for all $c \in \mathbb{R}$, the sublevel set $\{u \in E : f(u) \leqslant c\}$ is closed.*

It is clear that the product of an lsc function f by a positive scalar is lsc. A lower semicontinuous function is locally bounded below, as follows:

2.12 Proposition. *Let $f : E \to \mathbb{R}_\infty$ be lsc. If $x \in \operatorname{dom} f$, then for any $\varepsilon > 0$, there is a neighborhood V of x such that $f(u) > f(x) - \varepsilon \ \forall u \in V$. If $f(x) = \infty$, then for any $M \in \mathbb{R}$, there is a neighborhood V of x such that $f(u) > M \ \forall u \in V$.*

We omit the elementary proof of this, as well as of the following.

2.13 Proposition.

(a) *A positive linear combination of lsc functions is lsc.*

(b) *A function $f : E \to \mathbb{R}_\infty$ is lsc if and only if $\operatorname{epi} f$ is closed in $E \times \mathbb{R}$.*

(c) *The upper envelope of a family of lsc functions is lsc: if f_α is lsc for each index α, then the function f defined by $f(x) = \sup_\alpha f_\alpha(x)$ is lsc.*

A function f such that $-f$ is lower semicontinuous is called *upper semicontinuous*. Because we choose to emphasize minimization and convexity (rather than maximization and concavity), this property will play a lesser role.

Lower semicontinuity is a suitable replacement for continuity in Weierstrass's celebrated result concerning the existence of a minimum:

2.14 Exercise. Let E be a compact topological space and let $f : E \to \mathbb{R}_\infty$ be a proper lsc function. Prove that $\inf_E f$ is finite, and that f attains its minimum on E. □

If f is lsc, and if x_i is a sequence in E converging to x, then it follows easily that $f(x) \leqslant \liminf_{i\to\infty} f(x_i)$. In a metric setting, lower semicontinuity can be characterized in such sequential terms:

2.15 Proposition. *Let E be a metric space, and $f : E \to \mathbb{R}_\infty$. Then f is lsc if and only if for every $x \in E$ and $\ell \in \mathbb{R}$ we have*

$$\lim_{i\to\infty} x_i = x, \quad \lim_{i\to\infty} f(x_i) \leqslant \ell \implies f(x) \leqslant \ell .$$

2.3 Convex functions

The convexity of functions is destined to play an important role in later developments. The reader may as well see the definition immediately. Let $f : X \to \mathbb{R}_\infty$ be a given extended-valued function. We say that f is *convex* if it satisfies

$$f\big((1-t)x + ty\big) \leqslant (1-t)f(x) + t f(y), \quad x, y \in X, t \in (0,1).$$

In the inequality (and always in future), we interpret $t \times \infty$ as ∞ for $t > 0$. We seek to avoid the indeterminate expression $0 \times \infty$; this is why $t = 0$ and $t = 1$ were excluded above. With this convention in mind, and by iterating, it follows that f is convex if and only if

$$f\Big(\textstyle\sum_{i=1}^m t_i x_i\Big) \leqslant \sum_{i=1}^m t_i f(x_i)$$

for every convex combination of points $x_i \in X$, $m \geqslant 2$, $t_i \geqslant 0$, $\sum_{i=1}^m t_i = 1$.

A function g is called *concave* when the function $-g$ is convex.[1]

2.16 Exercise. Let Y be a vector space, $\Lambda \in L(X,Y)$, and y a point in Y. If the function $g : Y \to \mathbb{R}_\infty$ is convex, then the function $f(x) = g(\Lambda x + y)$ is convex. □

[1] Although X is a normed space throughout this chapter, it is clear that these basic definitions require only that X be a vector space.

2.17 Exercise. Let $f : X \to \mathbb{R}_\infty$ be positively homogeneous and subadditive. Prove that f is convex.[2] □

2.18 Exercise. Let Σ be a nonempty subset of X^*, and S a subset of X.

(a) The support function H_Σ (see Def. 2.10) is positively homogeneous and subadditive (and therefore, convex), as well as proper and lsc.

(b) The indicator function I_S is convex if and only if S is convex, lsc if and only if S is closed, and proper if and only if S is nonempty. □

On occasion it is useful to restrict attention to the values of f on a specified convex subset U of X. We say that f is convex on U if

$$f\big((1-t)x+ty\big) \leqslant (1-t)f(x)+tf(y), \quad x,y \in U,\ t \in (0,1).$$

It is clear that f is convex on U if and only if the function $f+I_U$ (that is, the function which coincides with f on U and which equals ∞ elsewhere) is convex.

We leave as an exercise the proof of the following.

2.19 Proposition. *Let $f : X \to \mathbb{R}_\infty$ be an extended-valued function. Then f is convex if and only if, for every segment $[x,y]$ in X, the function g defined by $g(t) = f((1-t)x+ty)$ is convex on $(0,1)$.*

The class of convex functions is closed under certain operations, in rather similar fashion to lower semicontinuous ones (see Prop. 2.13):

2.20 Proposition.

(a) *A positive linear combination of convex functions is convex.*

(b) *A function $f : E \to \mathbb{R}_\infty$ is convex if and only if* epi f *is a convex subset of $E \times \mathbb{R}$.*

(c) *The upper envelope of a family of convex functions is convex.*

Proof. We prove only the last assertion. Let the function f be defined as

$$f(x) = \sup_\alpha f_\alpha(x),$$

where, for each α, the function $f_\alpha : X \to \mathbb{R}_\infty$ is convex. Let $x,y \in X$, $t \in (0,1)$ be given. Then

$$\begin{aligned} f\big((1-t)x+ty\big) &= \sup_\alpha f_\alpha\big((1-t)x+ty\big) \leqslant \sup_\alpha \big\{(1-t)f_\alpha(x)+tf_\alpha(y)\big\} \\ &\leqslant (1-t)\sup_\alpha f_\alpha(x)+t\sup_\alpha f_\alpha(y) = (1-t)f(x)+tf(y). \end{aligned}$$ □

[2] We say that f is positively homogeneous if $f(tx) = tf(x)$ whenever t is a positive scalar. Subadditivity is the property $f(x+y) \leqslant f(x)+f(y)$.

2.21 Exercise. If $h\colon \mathbb{R}^n \to \mathbb{R}$ is convex, and if $\theta : \mathbb{R} \to \mathbb{R}$ is convex and increasing, prove that the function $f(x) = \theta(h(x))$ is convex. □

In the course of events, the reader will come to recognize the following recurrent theme: a convex function automatically benefits from a certain regularity, just because it is convex. Here is a first result of this type.

2.22 Proposition. *Let $f : X \to \mathbb{R}_\infty$ be convex, and $x \in \operatorname{dom} f$. Then the directional derivative $f'(x;v)$ exists for every $v \in X$, with values in $[-\infty,+\infty]$, and we have*

$$f'(x;v) = \inf_{t>0} \frac{f(x+tv)-f(x)}{t}.$$

Proof. It suffices to show that the function $g(t) = (f(x+tv)-f(x))/t$ is nondecreasing on the domain $t > 0$. Simply regrouping terms shows that, for $0 < s < t$, we have

$$g(s) \leqslant g(t) \iff f(x+sv) \leqslant (s/t)f(x+tv) + \big(1-(s/t)\big)f(x).$$

But this last inequality holds because f is convex. □

2.23 Example. The function $f : \mathbb{R} \to \mathbb{R}_\infty$ given by

$$f(x) = \begin{cases} -\sqrt{1-x^2} & \text{if } |x| \leqslant 1 \\ +\infty & \text{if } |x| > 1 \end{cases}$$

is convex, with $\operatorname{dom} f = [-1,1]$. We find $f'(1;-1) = -\infty$ and $f'(1;1) = +\infty$. Note that the derivative of f exists in the interior of $\operatorname{dom} f$, but becomes unbounded as one approaches the boundary of $\operatorname{dom} f$. □

2.24 Exercise. Let $f : X \to \mathbb{R}_\infty$ be convex. Prove the following assertions.

(a) f attains a minimum at $x \in \operatorname{dom} f$ if and only if $f'(x;v) \geqslant 0 \;\; \forall\, v \in X$.

(b) A finite local minimum of f is a global minimum. □

The necessary condition for a minimum expressed in Prop. 1.39 becomes a *sufficient* condition for optimality when the data are convex, as we now see.

2.25 Proposition. *Let $f : X \to \mathbb{R}$ be convex and differentiable, and let $A \subset X$ be convex. The point $x \in A$ minimizes f over A if and only if $-f'(x) \in N_A(x)$.*

Proof. We know the necessity of the condition $-f'(x) \in N_A(x)$ from Prop. 1.39; there remains to prove that this is a sufficient condition for x to be a solution of the optimization problem $\min_A f$ (when f and A are convex).

Let u be any point in A. Then $v := u-x$ belongs to $T_A(x)$, by Prop. 2.9, and we have (by Prop. 2.22)

$$f(u)-f(x) = f(x+v)-f(x) \geqslant \langle f'(x), v \rangle.$$

This last term is nonnegative, since $-f'(x) \in N_A(x)$, and since the normal cone is the polar of the tangent cone; it follows that $f(u) \geqslant f(x)$. □

Criteria for convexity. The following first and second order conditions given in terms of derivatives are useful for recognizing the convexity of a function.

2.26 Theorem. *Let U be an open convex set in X, and let $f : U \to \mathbb{R}$ be a function which is differentiable at each point of U.*

(a) *f is convex on U if and only if*

$$f(y)-f(x) \geqslant \big\langle f'(x), y-x \big\rangle, \quad x, y \in U. \tag{$*$}$$

(b) *If in addition f is twice continuously differentiable in U, then f is convex on U if and only if $f''(x)$ is positive semidefinite for every $x \in U$.*

Proof.

(a) Fix any two points $x, y \in U$. If f is convex, then

$$\big\langle f'(x), y-x \big\rangle = f'(x; y-x) \leqslant \frac{f\big(x+(y-x)\big)-f(x)}{1},$$

in light of Prop. 2.22, whence ($*$). Conversely, let us posit ($*$). For $0<t<1$, set $z=(1-t)x+ty$. Then $z \in U$, since U is convex. Invoking ($*$) reveals

$$f(y)-f(z) \geqslant \big\langle f'(z), y-z \big\rangle, \quad f(x)-f(z) \geqslant \big\langle f'(z), x-z \big\rangle.$$

We multiply these inequalities by t and $1-t$ respectively, and then add in order to obtain

$$f\big((1-t)x+ty\big) \leqslant (1-t)f(x)+tf(y),$$

which confirms the convexity of f.

(b) Let x, y be distinct points in U. Restricting f to an open segment containing x and y, and applying Lagrange's celebrated theorem (also known as the Taylor expansion with the Lagrange form of the remainder) on the real line, we obtain

$$f(y)-f(x)-\langle f'(x), y-x \rangle = (1/2)\langle f''(z)(y-x), y-x \rangle \text{ for some } z \in (x,y).$$

(Note that $f''(z)$ lies in $L_C(X, X^*)$.) If $f''(\cdot)$ is known to be positive semidefinite everywhere on U, the right side above is nonnegative, and the convexity of f follows from part (a) of the theorem.

For the converse, let us assume that f is convex. Then the left side above is nonnegative, by part (a). For $v \in X$ fixed, set $y = x + tv$, which lies in U when $t > 0$ is sufficiently small. We deduce

$$t^2 \langle f''(z)v, v \rangle \geqslant 0 \text{ for some } z \in (x, x+tv).$$

Dividing by t^2 and letting $t \downarrow 0$, we arrive at $\langle f''(x)v, v \rangle \geqslant 0$. Since $x \in U$ and $v \in X$ are arbitrary, it follows that $f''(\cdot)$ is positive semidefinite on U. □

Recall that for a C^2 function $f : \mathbb{R}^n \to \mathbb{R}$, the **Hessian matrix** refers to the $n \times n$ symmetric matrix $\nabla^2 f(x)$ defined by

$$\nabla^2 f(x) = \left[\frac{\partial^2 f}{\partial x_i \, \partial x_j}(x) \right] \quad \big(i, j = 1, 2, \ldots, n\big).$$

It induces the quadratic form corresponding to $f''(x)$. The well-known characterization of positive semidefinite matrices by means of eigenvalues leads to:

2.27 Corollary. *Let U be an open convex subset of $\mathbb{R}^n$, and let $f : U \to \mathbb{R}$ be C^2. Then f is convex on U if and only if, for every $x \in U$, all the eigenvalues of the Hessian matrix $\nabla^2 f(x)$ are nonnegative.*

A consequence of the corollary is that the convexity of a C^2 function f on an interval (a, b) in $\mathbb{R}$ is equivalent to the condition $f''(t) \geqslant 0 \;\forall t \in (a, b)$. This fact immediately implies the inequality used in the proof of Prop. 1.7, as the reader may show. We must emphasize, however, that the convexity of a function of several variables cannot be verified "one variable at a time."

2.28 Exercise. Prove that each of the following three functions is convex separately as a function of x (for each y) and as a function of y (for each x):

$$\exp(x+y), \;\; \exp(xy), \;\; \exp x + \exp y.$$

However, at least one of them fails to be convex on $\mathbb{R}^2$ (that is, jointly in (x, y)). Which ones are convex on $\mathbb{R}^2$? □

2.29 Exercise. Show that $x \mapsto \ln x$ is concave on the set $x > 0$. Deduce from this the inequality between the geometric and arithmetic means:

$$\sqrt[n]{a_1 \cdot a_2 \cdot \ldots \cdot a_n} \;\leqslant\; \frac{a_1 + a_2 + \cdots + a_n}{n} \quad (a_i > 0). \qquad \square$$

2.30 Example. Integral functionals, which are very important to us, are naturally extended-valued in many cases. Let $\Lambda : [0,1] \times \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ be continuous and bounded below. For $x \in X = \mathrm{AC}[0,1]$, we set

$$f(x) = \int_0^1 \Lambda\big(t, x(t), x'(t)\big)\, dt.$$

Under the given hypotheses, the composite function $t \mapsto \Lambda(t, x(t), x'(t))$ is measurable and bounded below, so that its (Lebesgue) integral is well defined, possibly as $+\infty$. When x is identically zero, or more generally, when x is a Lipschitz function, $x'(t)$ is bounded, and it follows that $f(x)$ is finite. Thus $f : X \to \mathbb{R}_\infty$ is proper. However, it is easy to construct an example in which x' is summable but unbounded ($x(t) = t^{1/2}$, say) and for which $f(x) = +\infty$ (take $\Lambda(t,x,v) = v^2$).

We claim that f is lsc. To see this, let x_i be a sequence in X converging to x, with $\lim_{i\to\infty} f(x_i) \leqslant \ell$. (We intend to use the criterion of Prop. 2.15.) Then x_i' converges in $L^1(0,1)$ to x', and $x_i(a) \to x(a)$. It follows easily from this that $x_i(t) \to x(t)$ for each t. Furthermore, there is a subsequence of x_i' (we do not relabel) that converges almost everywhere to x'. Then, by Fatou's lemma, we calculate

$$\begin{aligned} f(x) \;&=\; \int_0^1 \Lambda\big(t, x(t), x'(t)\big)\,dt \;=\; \int_0^1 \liminf_{i\to\infty} \Lambda\big(t, x_i(t), x_i'(t)\big)\,dt \\ &\leqslant\; \liminf_{i\to\infty} \int_0^1 \Lambda\big(t, x_i(t), x_i'(t)\big)\,dt \;=\; \lim_{i\to\infty} f(x_i) \;\leqslant\; \ell, \end{aligned}$$

whence the lower semicontinuity.

We ask the reader to show that if, for each $t \in [0,1]$, the function $(x,v) \mapsto \Lambda(t,x,v)$ is convex, then f is convex.

The integral functional f may be restricted to the subspace $\mathrm{AC}^p[0,1]$ of $\mathrm{AC}[0,1]$, an observation that has some importance when we consider calculating its directional derivatives. Suppose that Λ is C^1, and consider the case $p = \infty$ (thus, only Lipschitz functions x are involved). We suggest that the reader justify the formula

$$f'(x;y) \;=\; \int_0^1 \big\{ \Lambda_x(t) y(t) + \Lambda_v(t) y'(t) \big\}\,dt,$$

where $\Lambda_x(t)$ and $\Lambda_v(t)$ denote the partial derivatives of the function Λ evaluated at $(t, x(t), x'(t))$. (The proof involves switching a limit and an integral, a step for which Lebesgue's dominated convergence theorem can be invoked.) As we shall see later, deriving the formula is considerably more delicate when $1 \leqslant p < \infty$. □

The following functional property plays an important role in things to come.

2.31 Definition. (The Lipschitz property) *Let S be a subset of X, and let Y be a normed space. The function $g : S \to Y$ is said to be Lipschitz (of rank K, on S) if*

$$\| g(x) - g(y) \|_Y \;\leqslant\; K \| x - y \|_X \quad \forall x, y \in S.$$

It is said to be Lipschitz near x if, for some neighborhood V_x of x and some constant K_x, g is defined and Lipschitz on V_x of rank K_x. Finally, g is called **locally Lipschitz** *on an open set $U \subset X$ if it is Lipschitz near x for every $x \in U$.*

It is easy to see that a linear combination of functions that are Lipschitz on S is Lipschitz on S. Another useful fact is the following: if $f : X \to \mathbb{R}$ is continuously differentiable in a neighborhood of a point x, then f is Lipschitz near x; this is a consequence of the mean value theorem.

2.32 Exercise.

(a) Let $f : X \to \mathbb{R}$ be Lipschitz near each point x of a compact set C. Prove that f is Lipschitz on C.

(b) Let A be a nonempty subset of X. Show that the distance function d_A is Lipschitz of rank 1 on X.

(c) Let $\{ f_\alpha : S \to \mathbb{R} \}_\alpha$ be a family of functions, each of which is Lipschitz of rank K on S, such that the upper envelope $f(x) = \sup_\alpha f_\alpha(x)$ is finite-valued on S. Prove that f is Lipschitz of rank K on S. □

2.33 Proposition. *Let S be a nonempty subset of X, and let $f : S \to \mathbb{R}$ be Lipschitz of rank K. Then there exists $F : X \to \mathbb{R}$ extending f, and which is Lipschitz of rank K on X.*

Proof. The reader is asked to verify (with the help of part (c) of the preceding exercise) that $F(x) := \sup_{y \in S} \{ f(y) - K \| x - y \| \}$ does the trick. □

It turns out that convex functions have a natural predisposition to be continuous, even Lipschitz. This can only happen in the interior of the effective domain, of course. But even there, something more must be postulated. This can be seen by considering a discontinuous linear functional: it is convex, and its effective domain is the whole space, yet it is continuous at no point. The following shows that if a convex function f is "reasonable" at least at *one* point, then it is locally Lipschitz in the interior of $\operatorname{dom} f$.

2.34 Theorem. *Let $f : X \to \mathbb{R}_\infty$ be a convex function which admits a nonempty open set upon which f is bounded above. Then f is locally Lipschitz in the set* $\operatorname{int} \operatorname{dom} f$.

Proof. We require the following:

Lemma 1. *Let $f : X \to \mathbb{R}_\infty$ be convex, and let C be a convex set such that, for certain positive constants δ and N, we have $|f(x)| \leqslant N \;\; \forall x \in C + \delta B$. Then f is Lipschitz on C of rank $2N/\delta$.*

To prove the lemma, let us fix two distinct points x and y in C. The point z defined by $z = y + \delta(y - x)/\| y - x \|$ belongs to $C + \delta B$, and satisfies

$$y = \frac{\delta}{\delta + \| y - x \|}\, x + \frac{\| y - x \|}{\delta + \| y - x \|}\, z .$$

The convexity of f yields

$$f(y) \leqslant \frac{\delta}{\delta+\|y-x\|} f(x) + \frac{\|y-x\|}{\delta+\|y-x\|} f(z),$$

which implies

$$f(y)-f(x) \leqslant \frac{\big[f(z)-f(x)\big]\,\|y-x\|}{\delta+\|y-x\|} \leqslant \frac{2N}{\delta}\,\|y-x\|.$$

Since x and y are arbitrary points in C, this proves the lemma.

In view of Lemma 1, the theorem is now seen to follow from:

Lemma 2. *Let x_0 be a point such that, for certain numbers M and $\varepsilon > 0$, we have $f(x) \leqslant M \;\; \forall x \in B(x_0,\varepsilon)$. Then, for any $x \in \operatorname{int}\operatorname{dom} f$, there exists a neighborhood V of x and $N \geqslant 0$ such that $|f(y)| \leqslant N \;\; \forall y \in V$.*

Without loss of generality, we prove the lemma for $x_0 = 0$. Let $x \in \operatorname{int}\operatorname{dom} f$. There exists $r \in (0,1)$ such that $x/r \in \operatorname{dom} f$. Then

$$V := x+(1-r)B(0,\varepsilon) = B\big(x,(1-r)\,\varepsilon\big)$$

is a neighborhood of x. Every point u in this neighborhood can be expressed in the form $r(x/r)+(1-r)y$ for some $y \in B(0,\varepsilon)$, whence (by the convexity of f)

$$f(u) \leqslant rf(x/r)+(1-r)M =: M'.$$

Thus f is bounded above by M' on V. Now let $y \in V$. There exists $u \in V$ such that $(y+u)/2 = x$. Then we have

$$f(x) \leqslant (1/2)f(y)+(1/2)f(u) \leqslant (1/2)f(y)+M'/2,$$

which reveals that, on V, f is bounded below by $2f(x)-M'$. Since f is bounded both below and above on V, the required conclusion follows. □

We remark that Theorem 2.34 is false if "bounded above" is replaced by "bounded below." (Consider $f(x) = |\Lambda(x)|$, where Λ is a discontinuous linear functional.)

2.35 Corollary. *If X is finite dimensional, then any convex function $f : X \to \mathbb{R}_\infty$ is locally Lipschitz in the set $\operatorname{int}\operatorname{dom} f$.*

Proof. With no loss of generality, we may take $X = \mathbb{R}^n$. Let x_0 be any point in $\operatorname{int}\operatorname{dom} f$. By the theorem, it suffices to prove that f is bounded above in a neighborhood V of x_0. To see this, observe that, for some $r > 0$, we have

$$V := \operatorname{co}\big\{x_0 \pm re_i\big\}_i \subset \operatorname{dom} f,$$

where the e_i $(i = 1, 2, \ldots, n)$ are the canonical basis vectors in $\mathbb{R}^n$. Then, by the convexity of f, we deduce

$$f(y) \leqslant M := \max_i \big(|f(x_0 + re_i)| + |f(x_0 - re_i)| \big) \ \forall y \in V. \qquad \square$$

The gauge function. A convex set C for which $\operatorname{int} C \neq \emptyset$ is referred to as a *convex body*. For such a set, when $0 \in \operatorname{int} C$, the (Minkowski) **gauge** of C is the function g defined on X as follows:

$$g(x) = \inf \{ \lambda > 0 : x \in \lambda C \}.$$

It is clear that $g(x) \leqslant 1$ if $x \in C$. We claim that $g(x) \geqslant 1$ if $x \notin C$. For suppose the contrary: then there exists $\lambda \in (0,1)$ such that $x/\lambda \in C$. But then

$$x = (1-\lambda)0 + \lambda(x/\lambda)$$

expresses x as a convex combination of two points in the convex set C, whence $x \in C$, a contradiction. When $x \notin C$, then, roughly speaking, $g(x)$ is the factor by which the set C must be dilated in order to include the point x.

It is easy to see that the gauge of the unit ball is precisely the norm. The next result may be viewed as a generalization of this fact.

2.36 Theorem. *Let C be a convex subset of the normed space X for which $0 \in \operatorname{int} C$, and let g be its gauge. Then*

(a) *g has values in $[0, \infty)$.*

(b) $g(tx) = t\,g(x) \ \forall x \in X, t \geqslant 0.$

(c) $g(x+y) \leqslant g(x) + g(y) \ \forall x, y \in X.$

(d) *g is locally Lipschitz (and hence, continuous).*

(e) $\operatorname{int} C = \{ x : g(x) < 1 \} \subset C \subset \{ x : g(x) \leqslant 1 \} = \operatorname{cl} C.$

Proof. The first two assertions follow easily. If x/λ and y/μ belong to C, then the identity

$$\frac{x+y}{\lambda+\mu} = \frac{\lambda}{\lambda+\mu}\frac{x}{\lambda} + \frac{\mu}{\lambda+\mu}\frac{y}{\mu}$$

shows that $(x+y)/(\lambda+\mu)$ belongs to C. This observation yields the third assertion (subadditivity). A function which is positively homogeneous and subadditive (as is g) is convex. Further, we have $g(x) \leqslant 1$ on C. It follows from Theorem 2.34 that g is locally Lipschitz. The final assertion is left as an exercise. $\square$

2.4 Separation of convex sets

A set of the form $\{x \in X : \langle \zeta, x \rangle = c\}$, where $0 \neq \zeta \in X^*$ and c is a scalar, is referred to as a *hyperplane*. The sets

$$\{x \in X : \langle \zeta, x \rangle \leqslant c\} \text{ and } \{x \in X : \langle \zeta, x \rangle \geqslant c\}$$

are the associated *halfspaces*. Roughly speaking, we speak of two sets K_1 and K_2 as being *separated* if there is a hyperplane such that K_1 is contained in one of the associated halfspaces, and K_2 in the other.

The reader may be interested to know that the next result, which is known as the *separation theorem*, has often been nominated as the most important theorem in functional analysis.

2.37 Theorem. (Hahn-Banach separation) *Let K_1 and K_2 be nonempty, disjoint convex subsets of the normed space X. They can be separated in the two following cases:*

(a) *If K_1 is open, there exist $\zeta \in X^*$ and $\gamma \in \mathbb{R}$ such that*

$$\langle \zeta, x \rangle < \gamma \leqslant \langle \zeta, y \rangle \ \forall x \in K_1,\ y \in K_2.$$

(b) *If K_1 is compact and K_2 is closed, there exist $\zeta \in X^*$ and $\gamma_1, \gamma_2 \in \mathbb{R}$ such that*

$$\langle \zeta, x \rangle < \gamma_1 < \gamma_2 < \langle \zeta, y \rangle \ \forall x \in K_1,\ y \in K_2.$$

The second type of separation above is called *strict*.

Proof.

(a) Fix $\bar{x} \in K_1$ and $\bar{y} \in K_2$, and set $z = \bar{y} - \bar{x}$ and $C = K_1 - K_2 + z$. Then C is an open convex set containing 0; let p be its gauge function. Since $z \notin C$ (because K_1 and K_2 are disjoint), we have $p(z) \geqslant 1$. We prepare an appeal to Theorem 1.32, by defining $L = \mathbb{R}z$ and $\lambda(tz) = t$. We proceed to verify the hypotheses.

If $t \geqslant 0$, then $\lambda(tz) = t \leqslant t\,p(z) = p(tz)$. Consider now the case $t < 0$. Then we evidently have $\lambda(tz) = t \leqslant 0 \leqslant p(tz)$. Thus, we have $\lambda \leqslant p$ on L. Invoking the theorem, we deduce the existence of a linear functional ζ defined on X which extends λ (thus, ζ is nonzero) and which satisfies $\zeta \leqslant p$ on X. In particular, we have $\zeta \leqslant 1$ on C (a neighborhood of 0), which implies that ζ is continuous.

Now let $x \in K_1, y \in K_2$. Then $x - y + z \in C$. Bearing in mind that $\langle \zeta, z \rangle$ and $\langle \lambda, z \rangle$ are equal to 1, we calculate

$$\langle \zeta, x \rangle - \langle \zeta, y \rangle + 1 = \langle \zeta, x - y + z \rangle \leqslant p(x - y + z) < 1,$$

whence $\langle \zeta, x \rangle < \langle \zeta, y \rangle$. It follows that $\zeta(K_1)$ and $\zeta(K_2)$ are disjoint convex sets in $\mathbb{R}$ (that is, intervals), with $\zeta(K_1)$ lying to the left of $\zeta(K_2)$. Furthermore, $\zeta(K_1)$ is an *open* interval, by the lemma below. We set $\gamma = \sup \zeta(K_1)$ to obtain the desired conclusion.

Lemma. *Let ζ be a nonzero linear functional on X, and let V be an open subset of X. Then $\zeta(V)$ is an open subset of $\mathbb{R}$.*

To prove the lemma, take any point x such that $\zeta(x) = 1$; we may assume that V is nonempty. Let $\zeta(v)$ (for $v \in V$) be a point in $\zeta(V)$. Since V is open, there exists $\varepsilon > 0$ such that $v + tx \in V$ whenever $|t| < \varepsilon$. Then $\zeta(V)$ contains a neighborhood $(\zeta(v) - \varepsilon, \zeta(v) + \varepsilon)$ of $\zeta(v)$, proving the lemma.

(b) We now examine the second case of the theorem. A routine argument uses the compactness of K_1 in order to derive the existence of an open convex neighborhood V of 0 such that $K_1 + V$ and K_2 are disjoint.[3] We may now apply the first case of the theorem: there exists $\zeta \in X^*$ such that $\zeta(K_1 + V)$ is an interval lying to the left of $\zeta(K_2)$. But $\zeta(K_1)$ is a compact subset of the open interval $\zeta(K_1 + V)$, so that

$$\max \zeta(K_1) < \sup \zeta\big(K_1 + V\big) \leqslant \inf \zeta(K_2).$$

This implies the existence of γ_1, γ_2 as required. □

The conclusion of the separation theorem may fail if the sets K_1 and K_2 do not satisfy the extra hypotheses of either the first or the second case:

2.38 Exercise. Let $X = \ell^2$, and set

$$K_1 = \big\{ x = (x_1, x_2, \dots) \in X : x_i > 0 \ \forall i \big\}, \quad K_2 = \ell_c^\infty$$

(see Example 1.6). Show that these sets are disjoint convex subsets of X, but that there is no $\zeta \in X^*$ that satisfies $\langle \zeta, x \rangle < \langle \zeta, y \rangle \ \forall x \in K_1,\ y \in K_2$. □

The rest of this section derives some consequences of the separation theorem.

2.39 Theorem. *Let X be a normed space.*

(a) *X^* separates points in X:* $x, y \in X,\ x \neq y \implies \exists\, \zeta \in X^* : \langle \zeta, x \rangle \neq \langle \zeta, y \rangle$.

(b) *Let L be a subspace of X. If $x \notin \overline{L}$, then there exists $\zeta \in X^*$ such that $\langle \zeta, x \rangle = 1$ and $\zeta \equiv 0$ on L. Consequently, if the following implication holds:*

$$\zeta \in X^*,\ \langle \zeta, L \rangle = 0 \implies \zeta = 0,$$

then L is dense in X.

Proof. Left as an exercise. □

[3] Each $x \in K_1$ admits $r(x) > 0$ such that $B(x, 2r(x)) \subset X \setminus K_2$. Let $\{B(x_i, r(x_i))\}$ be a finite subcovering of K_1. Then we may take $V = \cap_i B^\circ(0, r(x_i))$.

Positive linear independence. A given set of vectors $\{\zeta_i : i = 1,2,\ldots,k\}$ in X^* is said to be *positively linearly independent* if the following implication holds:

$$\sum_{i=1}^{k} \lambda_i \zeta_i = 0,\ \lambda_i \geqslant 0 \implies \lambda_i = 0\ \forall i \in \{1,2,\ldots,k\}.$$

This property is related to the existence of a *decrease direction* v for the given set: an element v satisfying $\langle \zeta_i, v\rangle < 0\ \forall i$. This concept plays an important role in constrained optimization. The *nonexistence* of such a direction is equivalent to positive linear dependence, as we now see.

2.40 Exercise. The goal is to prove the following:

Proposition. *Let $\{\zeta_i : i = 1,2,\ldots,k\}$ be a finite subset in X^*. The following are equivalent*:

(a) *There is no $v \in X$ such that $\langle \zeta_i, v\rangle < 0\ \forall i \in \{1,2,\ldots,k\}$;*

(b) *The set $\{\zeta_i : i = 1,2,\ldots,k\}$ is positively linearly dependent*: *there exists a nonzero nonnegative vector $\gamma \in \mathbb{R}^k$ such that $\sum_1^k \gamma_i \zeta_i = 0$.*

Show first that (b) $\implies$ (a). Now suppose that (a) holds. Why does Theorem 2.37 apply to the sets

$$K_1 = \{y \in \mathbb{R}^k : y_i < 0\ \forall i \in \{1,2,\ldots,k\}\},$$
$$K_2 = \{(\langle \zeta_1, v\rangle, \langle \zeta_2, v\rangle, \ldots, \langle \zeta_k, v\rangle) : v \in X\}?$$

Use separation to deduce (b). □

2.41 Exercise. Let E be a vector space, and let $f_0, f_1, \ldots, f_n$ be linear functionals on E. Use the separation theorem to prove that the following are equivalent:

(a) There exists $(\lambda_1, \ldots, \lambda_n) \in \mathbb{R}^n$ such that $f_0 = \sum_{i=1}^n \lambda_i f_i$;

(b) There exists $M \geqslant 0$ such that $|f_0(x)| \leqslant M \max_{1 \leqslant i \leqslant n} |f_i(x)|\ \forall x \in E$;

(c) $x \in E,\ f_i(x) = 0\ \forall i \in \{1,\ldots,n\} \implies f_0(x) = 0$. □

Support functions redux. We defined earlier the support function H_Σ of a subset Σ of X^* (see Def. 2.10), a function that is defined on X. We now consider the support function of a nonempty subset S of X. This refers to the function $H_S : X^* \to \mathbb{R}_\infty$ defined on X^* by

$$H_S(\zeta) = \sup_{x \in S} \langle \zeta, x\rangle,\ \zeta \in X^*.$$

The support function transforms certain inclusions into functional inequalities:

2.42 Proposition. *Let C and D be nonempty subsets of X, with D closed and convex. Then $C \subset D$ if and only if $H_C \leqslant H_D$.*

Proof. It is clear that the inclusion implies the inequality. Let us prove the converse by contradiction, supposing therefore that the inequality holds, but that there is a point $\alpha \in C \setminus D$. We may separate the sets $\{\alpha\}$ and D according to the second case of Theorem 2.37: there exists $\zeta \in X^*$ and scalars γ_i such that

$$\langle \zeta, x \rangle < \gamma_1 < \gamma_2 < \langle \zeta, \alpha \rangle \;\; \forall x \in D.$$

(The order of the separation has been reversed, which simply corresponds to replacing ζ by $-\zeta$.) But this implies $H_D(\zeta) < H_C(\zeta)$, a contradiction. □

2.43 Corollary. *Closed convex subsets of X are characterized by their support functions: two closed convex sets coincide if and only if their support functions are equal.*

2.44 Exercise. Let D be a compact convex subset of $\mathbb{R}^n$ and $f : [a,b] \to D$ a measurable function. Prove that

$$\frac{1}{b-a} \int_a^b f(t)\, dt \;\in D.$$

□

2.45 Exercise. Let C and D be nonempty subsets of X, with D closed and convex. Let S be a nonempty bounded subset of X. Prove that

$$C \subset D \iff C + S \subset D + S.$$

Proceed to show that this equivalence is false in general, even in one dimension, if S is not bounded, or if D is not closed, or if D is not convex. □

Separation in finite dimensions. When the underlying space X is finite dimensional, the separation theorem can be refined somewhat, as we now show. Let D be a convex subset of $\mathbb{R}^n$. The geometric content of the following is that there is a hyperplane that passes through any boundary point of D in such a way that D lies entirely in one of the associated halfspaces. The reader is invited to observe that this does *not* correspond to either of the cases treated by Theorem 2.37.

2.46 Proposition. *Let D be a convex subset of $\mathbb{R}^n$, and let α be a point in its boundary: $\alpha \in \partial D$. Then $\alpha \in \partial \overline{D}$, and there exists a nonzero vector $\zeta \in \mathbb{R}^n$ such that*

$$\langle \zeta, x - \alpha \rangle \leqslant 0 \;\; \forall x \in \overline{D}.$$

Proof. By translating, we may do the proof *en français…non, pardon*, we may reduce to the case $\alpha = 0$. We proceed to prove that $0 \in \operatorname{int} \overline{D} \implies 0 \in \operatorname{int} D$; this will establish the first assertion of the proposition.

There exists $r > 0$ such that the points $\pm r e_i$ $(i = 1, 2, \dots, n)$ lie in $\overline{D}$, where the e_i are the canonical basis vectors of $\mathbb{R}^n$. Let these $2n$ points be denoted by

$x_1, x_2, \ldots, x_{2n}$. Then, for some $\delta > 0$, we have $2\delta B \subset \operatorname{co}\{x_j\}_j$. For each j, let y_j be a point in D satisfying $|y_j - x_j| < \delta$. Then (see Exer. 2.4)

$$\delta B + \delta B = 2\delta B \subset \operatorname{co}\big\{y_j + (x_j - y_j)\big\}_j \subset \operatorname{co}\{y_j\}_j + \delta B.$$

Since $\operatorname{co}\{y_j\}_j$ is compact and convex (Exer. 2.8), we deduce from this

$$\delta B \subset \operatorname{co}\{y_j\}_j \subset D$$

(see Exer. 2.45). It follows that $0 \in \operatorname{int} D$.

We now prove the second assertion of the proposition. We have $0 \in \partial \overline{D}$, whence there is a sequence x_i of points in $\mathbb{R}^n \backslash \overline{D}$ converging to 0. Let y_i be the closest point in $\overline{D}$ to x_i. It is clear that $y_i \to 0$.

Now set $\zeta_i = x_i - y_i \neq 0$, and extract a subsequence (if necessary) so that $\zeta_i / |\zeta_i|$ converges to a limit ζ. Fix any $x \in \overline{D}$. For $0 < t < 1$, the point $(1-t)y_i + tx$ belongs to $\overline{D}$, since $\overline{D}$ is convex by Theorem 2.2. Since y_i is closest to x_i in $\overline{D}$, we deduce

$$|(1-t)y_i + tx - x_i| \geqslant |y_i - x_i|.$$

Squaring both sides and simplifying, we find $2t\langle \zeta_i, x - y_i \rangle - t^2 |x - y_i|^2 \leqslant 0$. This leads to $\langle \zeta_i, x - y_i \rangle \leqslant 0$, and, in the limit, to $\langle \zeta, x \rangle \leqslant 0$, as required. □

As a corollary, we obtain a third case that can be added to the two of the separation theorem. Note that no openness or compactness hypotheses are made here concerning the sets to be separated; it is the finite dimensionality that compensates.

2.47 Corollary. *If X is finite dimensional, and if K_1 and K_2 are disjoint convex subsets of X, then there exists $\zeta \in X^*$ different from 0 such that*

$$\langle \zeta, x \rangle \leqslant \langle \zeta, y \rangle \;\; \forall x \in K_1,\; y \in K_2.$$

Proof. We may take $X = \mathbb{R}^n$ without loss of generality. Let $D = K_1 - K_2$, a set not containing 0. If 0 is not in the boundary of D, then, for $\varepsilon > 0$ sufficiently small, the sets $K_1 + \varepsilon B$ and K_2 are disjoint; the first case of Theorem 2.37 applies to this situation, and yields the result. If, to the contrary, we have $0 \in \partial D$, then the required conclusion is a direct consequence of Prop. 2.46. □

Existence of nonzero normals. The existence of nonzero normals is closely related to separation. The reader will recall (Prop. 2.9) that when C is convex, and when $\alpha \in C$, the normal cone to C at α is described by

$$N_C(\alpha) = \big\{\, \zeta \in X^* : \langle \zeta, x - \alpha \rangle \leqslant 0 \;\; \forall x \in C \,\big\}.$$

It follows that this normal cone is trivial (that is, reduces to $\{0\}$) when $\alpha \in \operatorname{int} C$. In certain applications, the question of whether $N_C(\alpha)$ is nontrivial for a point α in the

boundary of C is crucial. The following result summarizes the two principal cases in which this can be asserted directly.

2.48 Corollary. *Let C be a convex subset of X, and α a point in the boundary of C. Suppose that one of the two following conditions holds*: $\operatorname{int} C \neq \emptyset$, *or X is of finite dimension. Then $N_C(\alpha) \neq \{0\}$.*

Proof. Consider first the case in which X is finite dimensional. Then (for a suitable equivalent norm) it is isometric to $\mathbb{R}^n$ for some positive integer n (Theorem 1.22), via some isometry $T : X \to \mathbb{R}^n$. It follows that $T\alpha \in \partial(TC)$, so that, by Prop. 2.46, there is a nonzero $\zeta \in N_{TC}(T\alpha)$. The following lemma, whose simple proof is omitted, then yields the required assertion.

Lemma. *If $\zeta \in N_{TC}(T\alpha)$, then the formula $\Lambda x = \langle \zeta, Tx \rangle \quad \forall x \in X$ defines an element $\Lambda \in N_C(\alpha)$.*

(In fact, we have $\Lambda = T^*\zeta$, where T^* is the adjoint of T; see §1.4.)

Consider now the case $\operatorname{int} C \neq \emptyset$. We may then separate the open set $\operatorname{int} C$ (which is convex by Theorem 2.2) from the set $\{\alpha\}$, according to the first case of Theorem 2.37. There results an element ζ of X^* such that

$$\langle \zeta, x \rangle < \langle \zeta, \alpha \rangle \ \forall x \in \operatorname{int} C.$$

Note that ζ is necessarily nonzero. Since $\overline{C} = \overline{\operatorname{int} C}$ (by Theorem 2.2), the preceding inequality implies $\langle \zeta, x \rangle \leqslant \langle \zeta, \alpha \rangle \ \forall x \in \overline{C}$. Thus, ζ is a nonzero element of $N_C(\alpha)$. □

2.49 Exercise. The reader may feel a need to see an example of a closed convex set admitting a boundary point at which the normal cone is trivial; we give one now. Let $X = \ell^2$, and consider

$$C = \big\{ x \in X : |x_i| \leqslant 1/i \ \forall i \big\}.$$

Show that C is closed and convex, that $0 \in \partial C$, and that $N_C(0) = \{0\}$. □

Chapter 3
Weak topologies

It has been said that the existence of a minimum for an optimization problem such as $\min_A f$ may be a sensitive issue in an infinite dimensional space X, since the compactness of A may be difficult to secure. The compactness fails, notably, when A is the unit ball, and it is quite possible that a continuous linear functional may not attain a minimum over B_X. What if we were to change the topology on X, in an attempt to mitigate this lack of compactness?

We would presumably want to have *fewer* open sets: this is called *weakening* the topology. The reason behind this is simple: the fewer are the open sets in a topology, the more likely it becomes that a given set is compact. (Fewer open sets means there are fewer open coverings that must admit finite subcoverings.) On the other hand, the fewer open sets there are, the harder it is for a function defined on the space to be continuous (or lower semicontinuous), which is the other main factor in guaranteeing existence. The tension between these two contradictory pressures, the problem of finding the "right" topology that establishes a useful balance between them, is one of the great themes of functional analysis. In this chapter, we study the issue of weakening the topology (but in a way that remains consistent with the vector space structure).

3.1 Induced topologies

Let X be a vector space, and let Φ be a vector space of linear functionals defined on X. We require that Φ be *separating*: for every $x \neq 0$, there exists $\varphi \in \Phi$ such that $\varphi(x) \neq 0$. We shall study the topology[1] on X that is induced in a natural way by the family Φ.

[1] A topology on X is a collection of subsets of X that contains X itself and the empty set, and which is closed under taking arbitrary unions and finite intersections. The members of the collection are referred to as the open sets of the topology.

F. Clarke, *Functional Analysis, Calculus of Variations and Optimal Control*, Graduate Texts in Mathematics 264, DOI 10.1007/978-1-4471-4820-3_3,

Consider the subsets of X having the form

$$V(x,\varphi,r) = \{\, y \in X : |\varphi(y-x)| < r \,\}, \text{ where } x \in X,\ \varphi \in \Phi,\ r > 0.$$

The collection of all such sets generates a topology on X, the weakest (smallest, least open sets) topology on X that contains all members of the collection. We denote this topology $\sigma(X,\Phi)$.

The topology $\sigma(X,\Phi)$ is evidently the weakest topology on X rendering each element φ of Φ continuous, since φ is continuous if and only if $V(x,\varphi,r)$ is open for every $x \in X$ and $r > 0$. It is common to refer to $\sigma(X,\Phi)$ as *the weak topology induced by* Φ. The sets $V(x,\varphi,r)$ form a sub-base for $\sigma(X,\Phi)$, which means that the collection of finite intersections of such sets forms a base for $\sigma(X,\Phi)$. This, in turn, means that a set is open for the topology $\sigma(X,\Phi)$ if and only if it can be written as a union of sets each having the form

$$\bigcap_{i\in F} V(x_i,\varphi_i,r_i), \text{ where } F \text{ is finite, } x_i \in X,\ \varphi_i \in \Phi,\ r_i > 0.$$

The sets $V(x,\varphi,r)$ generate the topology $\sigma(X,\Phi)$ in this way. They play the role of the balls in the norm topology. Note that

$$V(x,\varphi,r) = x + V(0,\varphi,r),\ \ V(0,\varphi,tr) = tV(0,\varphi,r)\ (t > 0).$$

It follows from this that when U is an open set in $\sigma(X,\Phi)$, then $x+U$ and tU (for $t \neq 0$) are open as well. This allows us to say, for example, that if $x_i \to x$ in the topology $\sigma(X,\Phi)$, then $x_i - x \to 0$. Thus, the topology $\sigma(X,\Phi)$ is compatible with the vector space operations, as was the norm topology.

The following summarizes what we need to know about $\sigma(X,\Phi)$.

3.1 Theorem. (The induced topology)

(a) *The subsets of the form*

$$\bigcap_{i\in F} V(0,\varphi_i,r) \quad \big(\, r > 0,\ \varphi_i \in \Phi,\ F \text{ finite} \,\big)$$

are open neighborhoods of 0 *in* $\sigma(X,\Phi)$ *which form a local base at* 0: *every neighborhood of* 0 *in the topology* $\sigma(X,\Phi)$ *contains such a set.*

(b) *The topology* $\sigma(X,\Phi)$ *is Hausdorff, and (for the relevant product topologies) renders the mappings* $(x,u) \mapsto x+u$ *and* $(t,x) \mapsto tx$ *continuous. The operations of translation and dilation preserve open sets*:

$$U \in \sigma(X,\Phi),\ x \in X,\ t \neq 0 \implies U + x \in \sigma(X,\Phi) \text{ and } tU \in \sigma(X,\Phi).$$

The sets of the form

$$\bigcap_{i\in F} V(x,\varphi_i,r) \quad \big(\, r > 0,\ \varphi_i \in \Phi,\ F \text{ finite} \,\big)$$

constitute a base of open sets at x.

(c) *A linear functional is continuous on the topological space* $(X, \sigma(X,\Phi))$ *if and only if it belongs to* Φ.

(d) *A sequence* x_i *in* X *converges in the topology* $\sigma(X,\Phi)$ *to a limit* x *if and only if the sequence* $\varphi(x_i)$ *converges to* $\varphi(x)$ *(in* $\mathbb{R}$*) for every* $\varphi \in \Phi$.

(e) *If* Z *is a topological space, then the mapping* $\psi : Z \to (X, \sigma(X,\Phi))$ *is continuous if and only if the mapping* $\varphi \circ \psi : Z \to \mathbb{R}$ *is continuous for each* $\varphi \in \Phi$.

Proof.

(a) The assertion amounts to showing that any set of the form $\bigcap_{i\in F} V(x_i, \varphi_i, r_i)$, when it contains 0, also contains a set of the form $\bigcap_{i\in F} V(0, \varphi_i, r)$ for $r > 0$ sufficiently small. We have by hypothesis $|\varphi_i(x_i)| < r_i$ for each i; it suffices to choose $r < \min_{i\in F} \{ r_i - |\varphi_i(x_i)| \}$.

(b) We show that $\sigma(X,\Phi)$ is Hausdorff. Let x, z be distinct points in X. Since Φ is separating, there exists $\varphi \in \Phi$ such that $\varphi(x-z) \neq 0$. It follows that, for $r > 0$ sufficiently small, $V(x, \varphi, r)$ and $V(z, \varphi, r)$ are disjoint neighborhoods of x and z respectively.

We now verify the continuity of $(x,u) \mapsto x+u$ at $(0,0)$, for illustrative purposes. Let W be any neighborhood of 0 in the topology $\sigma(X,\Phi)$. It contains a set of the form $V = \bigcap_{i\in F} V(0, \varphi_i, r)$, by the above. If x, u belong to the open set $\bigcap_{i\in F} V(0, \varphi_i, r/2)$, we then have $x+u \in V \subset W$, confirming the continuity. The remaining assertions of part (b) are left as exercises.

(c) The elements of Φ are continuous for the topology $\sigma(X,\Phi)$ by construction. Now let f_0 be a linear functional that is continuous for $\sigma(X,\Phi)$. It follows that f_0 is bounded on a set of the form $V = \bigcap_{i\in F} V(0, \varphi_i, r)$. Then, for some $M > 0$,

$$x \in X,\ \varphi_i(x) = 0 \ \ \forall\, i \in F \implies |f_0(x)| \leqslant M.$$

But when this holds, we actually have

$$x \in X,\ \varphi_i(x) = 0 \ \ \forall\, i \in F \implies f_0(x) = 0,$$

since x can be replaced by tx for any $t \in \mathbb{R}$. Then, by Exer. 2.41, f_0 is a linear combination of the φ_i, $i \in F$. Since Φ is a vector space, we find $f_0 \in \Phi$, as claimed.

(d) It is clear that $x_i \to x$ implies $\varphi(x_i) \to \varphi(x)$, since each $\varphi \in \Phi$ is continuous. Now let x_i be a sequence such that $\varphi(x_i)$ converges to $\varphi(x)$ for each $\varphi \in \Phi$. Let W be any neighborhood of x; it contains a set of the form $V = \bigcap_{j\in F} V(x, \varphi_j, r)$. For each $j \in F$, there exists N_j such that

$$i \geqslant N_j \implies |\varphi_j(x_i - x)| < r.$$

Let us now set $N = \max_{j\in F} N_j$. Then it follows that

$$i \geqslant N \implies x_i \in \bigcap_{j \in F} V(x, \varphi_j, r) \subset W,$$

confirming that $x_i \to x$ for the topology $\sigma(X, \Phi)$.

(e) Left as an exception. □

Separation for the induced topology. The separation theorem 2.37, which plays such a central role in normed spaces, may be extended to the induced topology. As before, Φ is taken to be a separating vector space of linear functionals defined on the vector space X.

3.2 Theorem. *Let K_1 and K_2 be nonempty disjoint convex subsets of X.*

(a) *If K_1 is open for the topology $\sigma(X, \Phi)$, there exist $\zeta \in \Phi$ and $\gamma \in \mathbb{R}$ such that*

$$\langle \zeta, x \rangle < \gamma \leqslant \langle \zeta, y \rangle \;\; \forall x \in K_1,\; y \in K_2 .$$

(b) *If K_1 is compact and K_2 is closed for the topology $\sigma(X, \Phi)$, there exist $\zeta \in \Phi$ and $\gamma_1, \gamma_2 \in \mathbb{R}$ such that*

$$\langle \zeta, x \rangle < \gamma_1 < \gamma_2 < \langle \zeta, y \rangle \;\; \forall x \in K_1,\; y \in K_2 .$$

Proof. It is a matter of adapting the proof of Theorem 2.37. As before, we consider first the case in which K_1 is open. Fix $\bar{x} \in K_1$ and $\bar{y} \in K_2$, and set

$$z = \bar{y} - \bar{x}\,,\;\; C = K_1 - K_2 + z .$$

Then C is convex, open for the topology $\sigma(X, \Phi)$, and contains 0; let p be its gauge. The point z does not belong to C, since K_1 and K_2 are disjoint; therefore, we have $p(z) \geqslant 1$. Arguing as before, we proceed to invoke Theorem 1.32 with $L = \mathbb{R}z$ and $\lambda(tz) = t$. We deduce the existence of a linear functional ζ which extends λ and which satisfies $\Lambda \leqslant p$ on X. In particular, we have $\zeta \leqslant 1$ on C.

Lemma. *ζ is continuous for the topology $\sigma(X, \Phi)$.*

We observe that $\zeta \geqslant -1$ on $-C$, whence $|\,\zeta\,| \leqslant 1$ on the set $W := -C \cap C$, which is open for $\sigma(X, \Phi)$. Then, for any $\varepsilon > 0$, we have $|\,\zeta(\varepsilon W)| \leqslant \varepsilon$, which implies the continuity at 0, and hence everywhere, of the linear functional ζ for the topology $\sigma(X, \Phi)$. This proves the lemma.

Part (c) of Theorem 3.1 reveals that $\zeta \in \Phi$. It follows as before that $\zeta(K_1)$ and $\zeta(K_2)$ are intervals, with $\zeta(K_1)$ lying to the left of $\zeta(K_2)$, and that $\zeta(K_1)$ is open; we conclude by setting $\gamma = \sup \Lambda(K_1)$.

The reduction of the second case of the theorem to the first is carried out essentially as it was in the proof of Theorem 2.37; we omit the details. □

3.2 The weak topology of a normed space

An important case of the general scheme described above occurs when X is a normed space and $\Phi = X^*$. Note that X^* is separating, by Theorem 2.39. The resulting topology, designated $\sigma(X, X^*)$, is referred to as *the weak topology of* X.

The weak topology of X is certainly contained in the original norm topology, since the norm topology renders each $\varphi \in X^*$ continuous, and since $\sigma(X, X^*)$ is the weakest topology having that property. For that reason, the original norm topology is referred to as *the strong topology of* X; we denote it by $\sigma(X, \|\cdot\|_X)$.

3.3 Proposition. *Let X be finite dimensional. Then the weak topology on X coincides with the strong topology.*

Proof. It is enough to show that a ball $B(0,r)$ contains a weak neighborhood of 0, for this implies that every strongly open set is weakly open. For some n, there is an isometry $T : X \to \mathbb{R}^n$. Then $TB(0,r)$ is a neighborhood of 0 in $\mathbb{R}^n$, and so contains a set of the form $\{u \in \mathbb{R}^n : |e_i \bullet u| < \varepsilon \;\; \forall i\}$, where $\varepsilon > 0$ and the e_i are the canonical basis vectors in $\mathbb{R}^n$. It follows that $B(0,r)$ contains the set

$$\{x \in X : |e_i \bullet Tx| < \varepsilon \;\; \forall i\}.$$

But this set contains 0 and is weakly open, since the map $x \mapsto e_i \bullet Tx$ defines a continuous linear functional on X. □

The reader will quite rightly conclude from the above that, in finite dimensions, introducing the weak topology has not changed a thing. In the infinite dimensional case, however, the weak topology is *always* strictly weaker than the strong, as the next result shows.

3.4 Proposition. *Let X be a normed space of infinite dimension. Then every weak neighborhood of 0 contains a nontrivial subspace of X, and the weak topology is strictly contained in the strong topology.*

Proof. Let W be a weak neighborhood of 0. Then W contains a set of the form $\bigcap_{i=1}^{n} V(0, \varphi_i, r)$. Now the linear map

$$x \to \varphi(x) = \big(\, \langle \varphi_1, x \rangle, \langle \varphi_2, x \rangle, \dots, \langle \varphi_n, x \rangle \,\big) \in \mathbb{R}^n$$

cannot be injective, for otherwise X would be isomorphic as a vector space to the subspace $\varphi(X)$ of $\mathbb{R}^n$, whence $\dim X \leqslant n$. Thus, there exists $x_0 \neq 0$ such that $\varphi(x_0) = 0$. But then W contains the subspace $\mathbb{R}x_0$. Since the unit ball cannot contain a nontrivial subspace, it follows that it is not a weak neighborhood of 0, and that the weak topology is strictly weaker than the strong. □

3.5 Example. When one weakens a topology, one makes it more likely that a given sequence will be convergent. One expects, therefore, to find (in infinite dimensions) sequences that converge weakly and not strongly. We illustrate this now.

Let $1 < p < \infty$, and let e_i be the element of ℓ^p whose every term is 0 except for the i-th, which equals 1. Note that $\| e_i \|_{\ell^p} = 1$, so certainly, the sequence e_i does not converge to 0 in the norm topology. We claim that e_i converges weakly to 0, however.

In view of part (d) of Theorem 3.1, we may prove this by showing that $\langle \zeta, e_i \rangle \to 0$ for any $\zeta \in (\ell^p)^*$. Setting $q = p_*$, we know from Example 1.27 that ζ may be identified with an element $(\zeta_1, \zeta_2, \dots) \in \ell^q$, and we have $\langle \zeta, e_i \rangle = \zeta_i$. But clearly $\zeta_i \to 0$, since $\sum_{i \geqslant 1} |\zeta_i|^q < \infty$. □

There is one thing that the strong and weak topologies do agree on, and that is the following: which convex sets are closed?

3.6 Theorem. *Let C be a convex subset of X. Then the weak closure of C coincides with the strong closure of C. In particular, a convex subset of X is strongly closed if and only if it is weakly closed.*

Proof. Let F_s be the collection of all strongly closed sets A containing C. Then the closure of C with respect to the strong topology, denoted $\mathrm{cl}_s C$, coincides with $\bigcap_{A \in F_s} A$. Similarly, the closure of C with respect to the weak topology, denoted $\mathrm{cl}_w C$, coincides with $\bigcap_{A \in F_w} A$, where F_w is the collection of all weakly closed sets A containing C. Since a weakly closed set is strongly closed (this is clear from considering its complement), we have $F_w \subset F_s$, whence $\mathrm{cl}_s C \subset \mathrm{cl}_w C$.

To prove the opposite inclusion, we suppose that there is a point $x \in \mathrm{cl}_w C \setminus \mathrm{cl}_s C$, and we derive a contradiction. By the separation theorem 2.37, there exist $\zeta \in X^*$, $\gamma \in \mathbb{R}$ such that

$$\langle \zeta, x \rangle < \gamma < \langle \zeta, y \rangle \quad \forall y \in \mathrm{cl}_s C.$$

It follows that x does not belong to the set $S = \zeta^{-1}[\gamma, \infty)$. But S is weakly closed, since ζ is weakly continuous, and contains C. We deduce that $x \notin \mathrm{cl}_w C$, a contradiction which completes the proof. □

3.7 Corollary. *Let $f : X \to \mathbb{R}_\infty$ be a convex function. Then f is lsc if and only if f is weakly lsc (that is, lsc relative to the weak topology).*

Proof. The weak topology for $X \times \mathbb{R}$ is just the product topology $\sigma(X, X^*) \times \tau$, where τ denotes the usual topology on $\mathbb{R}$. The strong topology for $X \times \mathbb{R}$ is the product topology $\sigma(X, \|\cdot\|_X) \times \tau$. The lower semicontinuity of f on a topological space (X, σ) is equivalent to $\operatorname{epi} f$ being closed for the product topology $\sigma \times \tau$. But $\operatorname{epi} f$ is a convex set when f is a convex function. Consequently, weak and strong lower semicontinuity coincide in this case, by the theorem. □

3.8 Exercise. If $x_n \to x$ strongly, then $\|x\| = \lim_{n\to\infty} \|x_n\|$. If $x_n \to x$ weakly, then $\|x\| \leqslant \liminf_{n\to\infty} \|x_n\|$. (Note: Example 3.5 describes a case in which strict inequality holds.) □

The strong and weak topologies definitely disagree when it comes to the closure of sets which are not convex, as the following illustrates.

3.9 Example. Let X be an infinite dimensional normed space. Then the weak closure of the unit sphere in X is the closed unit ball.

The unit ball B is a weakly closed set, by Theorem 3.6, and it contains the unit sphere S. It is clear, then, that B contains the weak closure of S. To obtain the opposite inclusion, it suffices to prove that for any $x \in B$, for any weak neighborhood V of x, we have $V \cap S \neq \emptyset$. We may suppose in so doing that V is of the canonical form

$$\{\, u \in X : |\langle \zeta_i, u - x \rangle| < r \;\; \forall\, i \in F \,\},$$

where $\{\, \zeta_i : i \in F \,\}$ is a finite collection in X^* and $r > 0$. Arguing as we did in the proof of Prop. 3.4, we see that the infinite dimensionality of X implies the existence of $x_0 \neq 0$ such that $\langle \zeta_i, x_0 \rangle = 0 \;\; \forall\, i \in F$. Then, for every $\lambda \in \mathbb{R}$, the point $x + \lambda x_0$ belongs to V. The function $g(\lambda) := \lambda \mapsto \|x + \lambda x_0\|$ is continuous, with $g(0) \leqslant 1$ and $\lim_{\lambda\to\infty} g(\lambda) = \infty$. It follows that, for some $\lambda \geqslant 0$, we have $\|x + \lambda x_0\| = 1$; then $x + \lambda x_0 \in V \cap S$. □

3.10 Exercise. (Mazur's theorem) Let x_i be a sequence in X which converges weakly to x. We set

$$C = \operatorname{co}\{\, x_i : i = 1, 2 \ldots \,\},$$

the convex hull of the set $\{x_i : i \geqslant 1\}$. Prove the existence of a sequence y_i in C such that $\|y_i - x\|_X \to 0$. (Thus, there is a sequence of convex combinations of the terms x_i of the sequence that converges strongly to x.) □

We follow the usual practice in which a topological property on a normed space X, in the absence of any qualifier, is understood to be relative to the strong topology; thus "a closed set" means a strongly closed set.

3.3 The weak* topology

Another important special case of the general scheme studied in Theorem 3.1 arises when we take $X = Y^*$, where Y is a normed space, and where we take Φ to be the collection of all *evaluations* e_y at a point $y \in Y$:

$$e_y(x) = x(y), \;\; x \in X = Y^*.$$

Note that Φ is indeed a separating vector space of linear functionals on X, as required. The resulting topology on Y^* is called the weak* topology of Y^*, and is denoted by $\sigma(Y^*, Y)$. (Note: weak* is usually pronounced "weak star.")

A sequence ζ^j in Y^* converges to a limit ζ in the topology $\sigma(Y^*, Y)$ if and only if, for each $y \in Y$, the sequence of reals $\langle \zeta^j, y \rangle$ converges to $\langle \zeta, y \rangle$ for each y (this follows from Theorem 3.1 (d)). For this reason, the weak* topology is sometimes called the *topology of pointwise convergence*.

3.11 Example. Let ζ^j be the element of ℓ^∞ whose first j terms are 0, and whose terms are all equal to 1 thereafter. Evidently, the sequence ζ^j does not converge to 0, since $\|\zeta^j\| = 1 \ \forall j$. We claim that the sequence ζ^j converges in the weak* sense to 0. Note a convention here: strictly speaking, this assertion requires that ℓ^∞ be itself a dual space, whereas we only know it to be isometric to the dual space $(\ell^1)^*$; however, the underlying isometry is relegated to the status of a tacit understanding.

In view of Theorem 3.1(d), we need to verify that we have $\langle \zeta^j, y \rangle \to 0$ for any $y \in \ell^1$. But $\langle \zeta^j, y \rangle = \sum_{i>j} y_i$, which converges to 0 since $\sum_1^\infty |y_i| < \infty$. □

We remark that the sequence ζ^j in the exercise above does not converge weakly to 0. (This follows from Exer. 3.10.) Thus, we are dealing with three types of convergence in X^*: strong, weak, and weak*.

In the normed space X, (strongly) closed convex sets and weakly closed convex sets coincide, so there is only one kind of closed convex set, so to speak. The situation is generally different in X^*, since convex sets that are (strongly) closed in X^* (that is, for the topology of the dual norm) may fail to be closed for the weak* topology (see Exer. 8.16). This distinction makes itself felt in the following result, in which the hypothesis specifies weak* closedness, in order to obtain separation by an element of X, rather than by an element of the dual of X^*.

3.12 Proposition. *Let Σ be a nonempty weak* closed convex subset of X^*, and let $\zeta \in X^* \setminus \Sigma$. Then there exists $x \in X$ and $\gamma \in \mathbb{R}$ such that*

$$\langle \sigma, x \rangle < \gamma < \langle \zeta, x \rangle \ \forall \sigma \in \Sigma.$$

Proof. This is a special case of Theorem 3.2, since the continuous linear functionals for $\sigma(Y^*, Y)$ are precisely the evaluations, by Theorem 3.1, part (c). □

Let Σ be a nonempty subset of X^*. We have met its support function $H_\Sigma : X \to \mathbb{R}_\infty$, defined on X by $H_\Sigma(x) = \sup_{\sigma \in \Sigma} \langle \sigma, x \rangle$. When Σ is convex and weak* closed, this function characterizes Σ, as we now see.

3.13 Corollary. *Let Σ and Δ be nonempty subsets of X^*, with Δ weak* closed and convex. Then*

$$\Sigma \subset \Delta \iff H_\Sigma \leqslant H_\Delta.$$

Proof. We need only prove that if $H_\Sigma \leqslant H_\Delta$, then $\Sigma \subset \Delta$. Suppose to the contrary that there is a point $\zeta \in \Sigma \setminus \Delta$. Invoking Prop. 3.12, we find a point $x \in X$ and $\gamma \in \mathbb{R}$ such that

$$\langle \psi, x \rangle < \gamma < \langle \zeta, x \rangle \leqslant H_\Sigma(x) \;\; \forall \psi \in \Delta .$$

But then $H_\Delta(x) < H_\Sigma(x)$, a contradiction. □

The weak* topology possesses a precious compactness property:

3.14 Theorem. (Alaoglu) *Let V be a neighborhood of 0 in the normed space X. Then the set*

$$\Gamma = \{ \Lambda \in X^* : |\Lambda x| \leqslant 1 \text{ for all } x \in V \}$$

is weak compact in X^*.*

Proof. We set

$$P = \prod_{x \in X} \mathbb{R} = \mathbb{R}^X = \{ \zeta : X \to \mathbb{R} \}.$$

The product topology on P is (by definition) the weakest topology that renders continuous each *projection* $\pi_x : P \to \mathbb{R}$, defined by $\pi_x \zeta = \zeta(x)$ for $\zeta \in P$. The elements ζ of P which belong to Γ are those which satisfy the conditions

$$\pi_{x+y}\zeta = \pi_x \zeta + \pi_y \zeta , \quad \pi_{tx}\zeta = t\pi_x \zeta \;\; \forall x, y \in X, \; t \in \mathbb{R},$$

as well as $|\pi_v \zeta| \leqslant 1 \;\; \forall v \in V$. This reveals that Γ is closed in P.

Let $\zeta \in \Gamma$. For each $x \in X$, there exists $t_x > 0$ such that $t_x x \in V$. Thus, $\pi_x \zeta$ lies in $[-1/t_x, 1/t_x]$. We deduce from this:

$$\Gamma \subset \prod_{x \in X} [-1/t_x, 1/t_x] .$$

The set on the right is compact in the product topology by Tychonov's theorem, and it follows that Γ is compact in P.

A sub-base for the topology of P consists of the sets

$$\{ \zeta \in P : |\pi_x \zeta - t| < r \}, \;\; x \in X, \; t \in \mathbb{R}, \; r > 0.$$

The weak* topology $\sigma(X^*, X)$ on X^*, for its part, is generated by the sub-basic elements

$$\{ \zeta \in X^* : |\langle \zeta, x \rangle - t| < r \}, \;\; x \in X, \; t \in \mathbb{R}, \; r > 0.$$

We conclude therefore that the topology $\sigma(X^*, X)$ is nothing else but the trace topology of X^* viewed as a topological subspace of P. It follows that Γ is weak* compact in X^*. □

3.15 Corollary. *The dual ball $B_* = \{ \zeta \in X^* : \|\zeta\|_* \leqslant 1 \}$ is weak* compact. More generally, a subset of X^* which is bounded and weak* closed is weak* compact.*

Proof. The set Γ of the theorem reduces to B_* when $V = B(0,1)$, which yields the first assertion. It follows that rB_* is weak*compact for every $r > 0$, since dilation is a homeomorphism for induced topologies.

For the second assertion, let Σ be a bounded and weak*closed subset of X^*. Then, for some $r > 0$, Σ is a weak*closed subset of rB_*, which is weak*compact; thus, Σ inherits the weak*compactness. □

3.16 Exercise. Let x_i $(i = 0, 1, \ldots, n)$ be given points in X. Prove the existence of a point $\zeta_0 \in X^*$ which minimizes the function $\zeta \mapsto \langle \zeta, x_0 \rangle$ over the set

$$\{\, \zeta \in X^* : \|\zeta\|_* \leqslant 1, \langle \zeta, x_i \rangle = 0 \ (i = 1, 2, \ldots, n) \,\}. \qquad \square$$

3.4 Separable spaces

A normed space X is *separable* if it contains a countable set $S = \{x_i : i \geqslant 1\}$ which is dense: cl $S = X$. The reader will have met this property in topology, where one learns, for example, that the normed space $C(K)$ is separable whenever K is a compact subset of $\mathbb{R}^n$ (a consequence of the Weierstrass polynomial approximation theorem).

3.17 Example. Let us study the separability or otherwise of the spaces of sequences defined in Example 1.6. Consider the set S consisting of the elements of ℓ_c^∞ whose terms are all rational numbers. Then S is countable, as the countable union over $n \geqslant 1$ of those countably many points having at most n nonzero terms. It is easy to see that S is dense in ℓ_c^∞. Consider now $x \in \ell^1$, and let $\varepsilon > 0$. Then

$$\textstyle\sum_{i \geqslant 1} |x_i| < \infty \implies \lim_{n \to \infty} \sum_{i \geqslant n} |x_i| = 0.$$

Thus there exists N such that

$$\textstyle\sum_{i \geqslant N} |x_i| < \varepsilon.$$

For each of the finitely many terms x_i with $i < N$, there is a rational number y_i such that $|y_i - x_i| < \varepsilon/(N-1)$. Set $y_i = 0$ for $i \geqslant N$. Then $y \in S$, and $\|y - x\|_1 < 2\varepsilon$. We have proved that ℓ^1 is separable. A similar argument shows that every space ℓ^p when $1 \leqslant p < \infty$, is separable. We claim, however, that ℓ^∞ *fails* to be separable. Let us prove this by contradiction, by supposing that there is a countable dense subset $S = \{x^i\}$ of $X = \ell^\infty$. Then the balls $B(x^i, 1/3)$ cover X. Consider now the elements $d = (d_1, d_2, d_3 \ldots)$ of ℓ^∞ which correspond to the binary expansions $.d_1 d_2 d_3 \ldots$ of the real numbers in $(0,1)$. Note that any two distinct elements generated this way are of distance 1 from one another. There are uncountably many such elements, so there is a ball $B(x^i, 1/3)$ containing two distinct ones. But then (by the triangle inequality) the distance between the two does not exceed $2/3$: contradiction. □

3.18 Exercise. Prove that c and c_0 are separable. □

We shall prove in Chapter 6 that the Lebesgue spaces $L^p(\Omega)$ defined in Example 1.9 are separable, provided $1 \leqslant p < \infty$.

It can be shown that a linear subspace of a separable space is separable; this is simply a special case of a more general fact regarding metric spaces. It is clear that the finite Cartesian product of separable spaces is itself separable. Since isometry preserves the density of a set, and since the space $\mathrm{AC}^p[a,b]$ is isometric to $\mathbb{R} \times L^p(a,b)$, it follows that $\mathrm{AC}^p[a,b]$ is separable when $1 \leqslant p < \infty$.

Concerning the dual space, it is the separability of the child that is inherited by the parent:

3.19 Theorem. *If the dual of a normed space X is separable, then X is separable.*

Proof. Let ζ_n be a dense sequence in X^*. By definition of $\|\zeta_n\|_*$, there exist $x_n \in B$ such that $\langle \zeta_n, x_n \rangle \geqslant \|\zeta_n\|_*/2$. Let L_0 be the vector space over the rationals $\mathbb{Q}$ generated by these points x_n:

$$L_0 = \operatorname{vect}_{\mathbb{Q}}\{x_1\} \cup \operatorname{vect}_{\mathbb{Q}}\{x_1, x_2\} \cup \ldots$$

We observe that L_0 is countable. Let L be the vector space over $\mathbb{R}$ generated by the points x_n. Then L_0 is dense in L, and in order to prove the theorem, it suffices to show that L is dense in X. We accomplish this by proving the implication (see Theorem 2.39)

$$\zeta \in X^*,\ \ \langle \zeta, x \rangle = 0 \ \forall x \in L \implies \zeta = 0.$$

Given $\varepsilon > 0$, there exists n such that $\|\zeta - \zeta_n\|_* < \varepsilon$. We have

$$\|\zeta_n\|_*/2 \leqslant \langle \zeta_n, x_n \rangle = \langle \zeta_n - \zeta, x_n \rangle + \langle \zeta, x_n \rangle = \langle \zeta_n - \zeta, x_n \rangle < \varepsilon.$$

It follows that $\|\zeta\|_* \leqslant \|\zeta - \zeta_n\|_* + \|\zeta_n\|_* \leqslant 3\varepsilon$, whence $\zeta = 0$. □

That the converse of Theorem 3.19 fails may be illustrated with the case $X = \ell^1$. As we know, X is a separable normed space. Its dual, however, is isometric to ℓ^∞, and therefore fails to be separable.

An important consequence of separability is the following.

3.20 Theorem. *If the dual of a normed space X is separable, and if S is a bounded subset of X, then the weak topology of X, when restricted to S, is metrizable.*

Proof. The goal here is to exhibit a metric ρ on S inducing a topology which is the same (has the same open sets) as the trace (or restriction) of the weak topology $\sigma(X, X^*)$ to S. Let $\{\zeta_i\}$ be a countable dense set in the dual; then the function defined by

$$\rho(x,y) = \sum_{i=1}^{\infty} \frac{|\langle \zeta_i, x-y \rangle|}{1+|\langle \zeta_i, x-y\rangle|} 2^{-i}$$

will serve the purpose (there are other possible choices). It is routine to verify that ρ is a metric on S (or indeed, on X). Let

$$B_\rho(x,R) = \{ y \in X : \rho(x,y) \leqslant R \}.$$

Then the equivalence of the resulting metric topology on S to the restriction to S of the weak topology amounts to showing two things:

(a) Each set of the form $B_\rho(x,R)$ contains an element $\bigcap_{i\in F} V(x,\varphi_i,r)$ of the canonical base for the weak topology (see §3.1);

(b) Each element $V(x,\varphi,r)$ of the canonical sub-base for the weak topology contains a set of the form $B_\rho(x,R) \cap S$.

The details are not of the greatest interest, and we omit them. We remark, however, that it is the second step above that requires the boundedness of S. □

Note: The theorem does *not* say that the weak topology on X is metrizable (when X^* is separable); indeed, this is never the case in infinite dimensions, as is shown in Exer. 8.43.

As regards the weak* topology, the analogous result is the following:

3.21 Theorem. *If the normed space X is separable, then the weak* topology of X^*, when restricted to a bounded set, is metrizable.*

The motivating factor behind our interest in metrizability is the possibility of invoking *sequential compactness* in certain contexts to come. The reader will recall that in general topological spaces, the properties of compactness and sequential compactness differ; in metric topologies, however, they coincide. Since (in infinite dimensions) the weak topologies are not metric ones, the extraction of a convergent subsequence must therefore be justified, even if the sequence lies in a compact set. This is what a result such as Theorem 3.21 allows us to do.

The following exercise serves to illustrate these considerations.

3.22 Exercise. Let ζ_i be a bounded sequence of continuous linear functionals on a separable normed space X. Prove the existence of a subsequence ζ_{i_j} such that, for each $x \in X$, we have $\langle \zeta_{i_j}, x \rangle \to \langle \zeta, x \rangle$ as $j \to \infty$. □

Chapter 4
Convex analysis

The phrase *convex analysis* refers to a body of generalized calculus that can be developed for convex functions and sets. This topic, whose applications are widespread, is the subject of the chapter. The central element of the theory is the *subdifferential*, a construct which plays a role similar to that of the derivative. The operation of *conjugacy* as it applies to convex functions will also be important, as well as *polarity* of sets.

4.1 Subdifferential calculus

Let $f : X \to \mathbb{R}_\infty$ be a given function, where X is a normed space, and let x be a point in $\operatorname{dom} f$. An element ζ of X^* is called a **subgradient** of f at x (in the sense of convex analysis) if it satisfies the following *subgradient inequality*:

$$f(y) - f(x) \geqslant \langle \zeta, y - x \rangle, \quad y \in X.$$

A function is called *affine* when it differs by a constant from a linear functional. Thus, an affine function g has the form $g(y) = \langle \zeta, y \rangle + c$; the linear functional ζ is called the *slope* of g. When the subgradient inequality above holds, the affine function

$$y \mapsto f(x) + \langle \zeta, y - x \rangle$$

is said to *support* f at x; this means that it lies everywhere below f, and that equality holds at x. In geometric terms, we may formulate the situation as follows: the hyperplane $\{(y,r) : r - f(x) - \langle \zeta, y - x \rangle = 0\}$ in the product space $X \times \mathbb{R}$ passes through the point $(x, f(x))$, and the set $\operatorname{epi} f$ lies in the upper associated halfspace (see p. 41 for the terminology). We refer to this as a *supporting hyperplane*.

The set of all subgradients of f at x is denoted by $\partial f(x)$, and referred to as the **subdifferential** of f at x. It follows from the definition that the subdifferential $\partial f(x)$ is a convex set which is closed for the weak* topology, since, for each y, the

F. Clarke, *Functional Analysis, Calculus of Variations and Optimal Control*, Graduate Texts in Mathematics 264, DOI 10.1007/978-1-4471-4820-3_4,

set of ζ satisfying the subgradient inequality is weak* closed and convex.[1] The map $x \mapsto \partial f(x)$ is set-valued: its values are subsets of X^*. We use the term *multifunction* in such a case: ∂f is a multifunction from X to X^*.

4.1 Example. We illustrate the geometry of subgradients with the help of Fig. 4.1, which we think of as depicting the epigraph of a convex function $f : \mathbb{R}^n \to \mathbb{R}_\infty$.

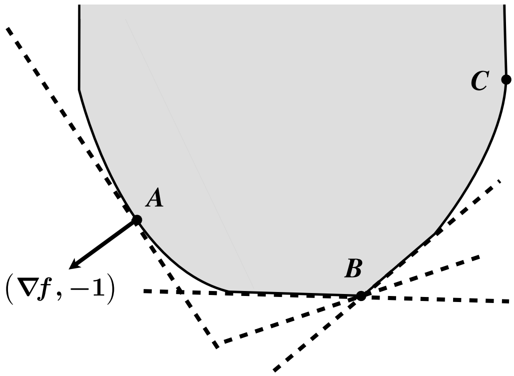


Fig. 4.1
The epigraph of a convex function, and some supporting hyperplanes.

The function is smooth near the point A on the boundary of its epigraph; let this point be $(x_1, f(x_1))$. There is a unique affine function $y = \langle \zeta, x \rangle + c$ that supports f at the point x_1; its slope ζ is given by $\nabla f(x_1)$. The vector $(\nabla f(x_1), -1)$ is orthogonal to the corresponding supporting hyperplane, and generates the normal cone to epi f at $(x_1, f(x_1))$ (here, a ray).

At the point B, which we take to be $(x_2, f(x_2))$, the function f has a corner, and there are infinitely many affine functions supporting f at x_2; the set of all their slopes constitutes $\partial f(x_2)$. There is a supporting hyperplane to epi f at the point C as well, but it is vertical, and therefore does not define a subgradient (it fails to correspond to the graph of an affine function of x). The subdifferential of f is empty at the corresponding value of x. □

4.2 Exercise. (subdifferential of the norm) Let f be the function $f(x) = \|x\|$.

(a) Prove that $\partial f(0)$ is the closed unit ball in X^*.

(b) Let $\zeta \in \partial f(x)$, where $x \neq 0$. Prove that $\langle \zeta, x \rangle = \|x\|$ and $\|\zeta\|_* = 1$. □

It is clear from the definition of subgradient that f attains a minimum at x if and only if $0 \in \partial f(x)$. This version of Fermat's rule is but the first of several ways in which the reader will detect a kinship between the subdifferential and the derivative. The following provides another example.

---

[1] In speaking, the subdifferential ∂f is often pronounced "dee eff" or "curly dee eff".

4.3 Proposition. *Let* $f : X \to \mathbb{R}_\infty$ *be a convex function, and* $x \in \operatorname{dom} f$*. Then*

$$\partial f(x) = \big\{\, \zeta \in X^* : f'(x;v) \geqslant \langle\, \zeta, v\rangle \;\; \forall v \in X \,\big\}.$$

Proof. We recall that a convex function admits directional derivatives, as showed in Prop. 2.22. If $\zeta \in \partial f(x)$, then we have

$$f(x+tv) - f(x) \geqslant \langle\, \zeta, tv\rangle \;\; \forall v \in X,\; t > 0,$$

by the subgradient inequality. It follows that $f'(x;v) \geqslant \langle\, \zeta, v\rangle \;\; \forall v$. Conversely, if this last condition holds, then (by Prop. 2.22) we have

$$f(x+v) - f(x) \geqslant \inf_{t>0} \frac{f(x+tv) - f(x)}{t} \geqslant \langle\, \zeta, v\rangle \;\; \forall v \in X,$$

which implies $\zeta \in \partial f(x)$. □

It is a consequence of the proposition above that if f is differentiable at x, then $\partial f(x) = \{\, f'(x)\}$. The reduction of $\partial f(x)$ to a singleton, however, is more closely linked to a weaker type of derivative, one that we proceed to introduce.

The Gâteaux derivative. Let $F : X \to Y$ be a function between two normed spaces. We say that F is *Gâteaux differentiable* at x if the directional derivative $F'(x;v)$ exists for all $v \in X$, and if there exists $\Lambda \in L_C(X,Y)$ such that

$$F'(x;v) = \langle \Lambda, v\rangle \;\; \forall v \in X.$$

It follows that the element Λ is unique; it is denoted $F'_G(x)$ and referred to as the Gâteaux derivative. It corresponds to a weaker concept of differentiability than the Fréchet derivative $F'(x)$ that we met in §1.4. In fact, Gâteaux differentiability at x does not even imply continuity of F at x. When F is Fréchet differentiable at x, then F is Gâteaux differentiable at x, and $F'_G(x) = F'(x)$. We stress that the unqualified word "differentiable" always refers to the usual (Fréchet) derivative.

The following is a direct consequence of Prop. 4.3.

4.4 Corollary. *Let* $f : X \to \mathbb{R}_\infty$ *be convex, with* $x \in \operatorname{dom} f$*. If* f *is Gâteaux differentiable at* x*, then* $\partial f(x) = \{\, f'_G(x)\}$*.*

A characteristic of convex analysis that distinguishes it from classical differential analysis is the close link that it establishes between sets and functions. The following records an important (yet simple) example of this.

4.5 Exercise. Let $x \in S$, where S is a convex subset of X. Prove that $\partial I_S(x) = N_S(x)$; that is, the subdifferential of the indicator function is the normal cone. □

For the convex function f of Example 2.23, the reader may check that $\partial f(-1)$ and $\partial f(1)$ are empty, and that $\partial f(x)$ is a singleton when $-1 < x < +1$. It is possible for the subdifferential of a convex function to be empty at points in the interior of its effective domain: if Λ is a discontinuous linear functional, then $\partial \Lambda(x)$ is empty for each x, as the reader may care to show. This phenomenon does not occur at points of continuity, however, as we now see.

4.6 Proposition. *Let $f : X \to \mathbb{R}_\infty$ be a convex function, and let $x \in \operatorname{dom} f$ be a point of continuity of f. Then $\partial f(x)$ is nonempty and weak* compact. If f is Lipschitz of rank K in a neighborhood of x, then $\partial f(x) \subset K B_*$.*

Proof. The continuity at x implies that $\operatorname{int} \operatorname{epi} f \neq \emptyset$. We separate $\operatorname{int} \operatorname{epi} f$ and the point $(x, f(x))$ (using the first case of Theorem 2.37) to deduce the existence of $\zeta \in X^*$ and $\lambda \in \mathbb{R}$ such that

$$\langle \zeta, y \rangle + \lambda r < \langle \zeta, x \rangle + \lambda f(x) \quad \forall (y, r) \in \operatorname{int} \operatorname{epi} f.$$

It follows that $\lambda < 0$; we may therefore normalize by taking $\lambda = -1$. Since

$$\operatorname{epi} f \subset \operatorname{cl} \operatorname{epi} f = \operatorname{cl} \big(\operatorname{int} \operatorname{epi} f \big)$$

(by Theorem 2.2), the separation inequality implies

$$\langle \zeta, y \rangle - r \leqslant \langle \zeta, x \rangle - f(x) \quad \forall (y, r) \in \operatorname{epi} f.$$

This reveals that $\partial f(x)$ contains ζ, and is therefore nonempty.

Theorem 2.34 asserts that f is Lipschitz on some neighborhood of x. Let K be a Lipschitz constant for f on $B(x, r)$, and let ζ be any element of $\partial f(x)$. The subgradient inequality yields

$$\langle \zeta, y - x \rangle \leqslant f(y) - f(x) \leqslant |f(y) - f(x)| \leqslant K |y - x| \quad \forall y \in B(x, r).$$

Putting $y = x + rv$, where v is a unit vector, leads to

$$\langle \zeta, v \rangle \leqslant K |v| \quad \forall v \in B,$$

whence $|\zeta| \leqslant K$. Thus, $\partial f(x)$ is bounded, and weak* compact by Cor. 3.15. □

4.7 Corollary. *Let $f : \mathbb{R}^n \to \mathbb{R}$ be a convex function. Then for any $x \in \mathbb{R}^n$, $\partial f(x)$ is a nonempty convex compact set.*

Proof. This follows from Cor. 2.35. □

Recall that we have agreed to identify the dual of $\mathbb{R}^n$ with $\mathbb{R}^n$ itself. Thus, for a function f defined on $\mathbb{R}^n$, the subdifferential $\partial f(x)$ is viewed as a subset of $\mathbb{R}^n$. The very existence of a subgradient is the key to the proof of the next result, a well-known inequality.

4.8 Corollary. (Jensen's inequality) *Let $\varphi : \mathbb{R}^k \to \mathbb{R}$ be convex. Then, for any summable function $g : (0,1) \to \mathbb{R}^k$, we have*

$$\varphi\Big(\int_0^1 g(t)\,dt\Big) \leqslant \int_0^1 \varphi\big(g(t)\big)\,dt.$$

Proof. Let us define a point in $\mathbb{R}^k$ by

$$\bar g := \int_0^1 g(t)\,dt,$$

the integral being understood in the vector sense. By Cor. 4.7, $\partial\varphi(\bar g)$ contains an element ζ. Then, by definition, we have $\varphi(y) - \varphi(\bar g) \geqslant \langle \zeta, y - \bar g \rangle \ \forall y \in \mathbb{R}^k$. Substituting $y = g(t)$ and integrating over $[0,1]$, we obtain the stated inequality. □

4.9 Exercise.

(a) Modify the statement of Jensen's inequality appropriately when the underlying interval is $[a,b]$ rather than $[0,1]$.

(b) Formulate and prove Jensen's inequality in several dimensions, when the function g belongs to $L^1(\Omega, \mathbb{R}^k)$, Ω being a bounded open subset of $\mathbb{R}^n$. □

The appealing calculus formula $\partial(f+g)(x) = \partial f(x) + \partial g(x)$ turns out to be true under mild hypotheses, as we see below. Note that *some* hypothesis is certainly required in order to assert such a formula, since it fails when we take $f = \Lambda$ and $g = -\Lambda$, where Λ is a discontinuous linear functional.

4.10 Theorem. (Subdifferential of the sum) *Let $f, g : X \to \mathbb{R}_\infty$ be convex functions which admit a point in $\operatorname{dom} f \cap \operatorname{dom} g$ at which f is continuous. Then we have*

$$\partial\big(f+g\big)(x) = \partial f(x) + \partial g(x) \ \ \forall x \in \operatorname{dom} f \cap \operatorname{dom} g.$$

Proof. That the left side above contains the right follows directly from the definition of subdifferential. Now let ζ belong to $\partial(f+g)(x)$; we must show that ζ belongs to the right side. We may (and do) reduce to the case $x = 0$, $f(0) = g(0) = 0$. By hypothesis, there is a point $\bar x$ in $\operatorname{dom} f \cap \operatorname{dom} g$ at which f is continuous. Then the subsets of $X \times \mathbb{R}$ defined by

$$C = \operatorname{int}\operatorname{epi} f, \ \ D = \big\{ (w,t) : \langle \zeta, w \rangle - g(w) \geqslant t \big\}$$

are nonempty as a consequence of the existence of $\bar x$. They are also convex, and disjoint as a result of the subgradient inequality for ζ. By Theorem 2.37, C and D can be separated: there exist $\xi \in X^*$ and $\lambda \in \mathbb{R}$ such that

$$\langle \xi, w \rangle + \lambda t < \langle \xi, u \rangle + \lambda s \ \forall (w,t) \in D, \ \ \forall (u,s) \in \operatorname{int}\operatorname{epi} f.$$

It follows that $\lambda > 0$; we can normalize by taking $\lambda = 1$. Since (by Theorem 2.2) we have $\operatorname{epi} f \subset \operatorname{cl}\operatorname{epi} f = \operatorname{cl}(\operatorname{int}\operatorname{epi} f)$, we deduce

$$\langle \xi, w \rangle + t \leqslant \langle \xi, u \rangle + s \;\; \forall (w,t) \in D, \;\; \forall (u,s) \in \operatorname{epi} f.$$

Taking $(w,t) = (0,0)$, this implies $-\xi \in \partial f(0)$. Taking $(u,s) = (0,0)$ leads to the conclusion $\xi + \zeta \in \partial g(0)$. Thus, $\zeta \in \partial f(0) + \partial g(0)$. □

4.11 Exercise. Let C and D be convex subsets of X such that $(\operatorname{int} C) \cap D \neq \emptyset$. Let $x \in C \cap D$. Then $N_{C \cap D}(x) = N_C(x) + N_D(x)$. □

The following Fermat-type result is related to that of Prop. 2.25; f is now nondifferentiable (which means "not necessarily differentiable"), but convex, and the necessary condition turns out to be sufficient as well.

4.12 Proposition. *Let $f : X \to \mathbb{R}$ be a continuous convex function, A a convex subset of X, and x a point in A. Then the following are equivalent:*

(a) *x minimizes f over the set A.*

(b) *$-\partial f(x) \cap N_A(x) \neq \emptyset$; or equivalently, $0 \in \partial f(x) + N_A(x)$.*

Proof. If (a) holds, then x minimizes $u \mapsto f(u) + I_A(u)$. Thus $0 \in \partial(f + I_A)(x)$, by Fermat's rule. But we have

$$\partial\big(f + I_A\big)(x) = \partial f(x) + \partial I_A(x) = \partial f(x) + N_A(x),$$

by Theorem 4.10 and Exer. 4.5; thus, (b) holds. Conversely, if (b) holds, then we have $0 \in \partial(f + I_A)(x)$, which implies (a). □

Recall that the adjoint of a linear application T is denoted T^* (see p. 22). It plays a role in the next result, a calculus rule for a composition.

4.13 Theorem. *Let Y be a normed space, $T \in L_C(X,Y)$, and let $g : Y \to \mathbb{R}_\infty$ be a convex function. Let there be a point x_0 in X such that g is continuous at Tx_0. Then the function $f(x) = g(Tx)$ is convex, and we have*

$$\partial f(x) = T^* \partial g(Tx) \;\; \forall x \in \operatorname{dom} f.$$

Note: the meaning of this formula is that, given any $\zeta \in \partial f(x)$, there exists γ in $\partial g(Tx)$ such that

$$\langle \zeta, v \rangle = \langle T^* \gamma, v \rangle = \langle \gamma, Tv \rangle \;\; \forall v \in X,$$

and that, conversely, any element ζ of the form $T^*\gamma$, where $\gamma \in \partial g(Tx)$, belongs to $\partial f(x)$.

Proof. It is easily verified that f is convex, and that any element of $T^*\partial g(Tx)$ lies in $\partial f(x)$. We now prove the opposite inclusion. Let $\varphi : X\times Y \to \mathbb{R}_\infty$ be defined by

$$\varphi(x,y) = g(y) + I_{\operatorname{gr} T}(x,y),$$

where gr T is the graph of T: the set $\{(x,Tx) \in X\times Y : x\in X\}$. It follows from the definition of subgradient that

$$\zeta \in \partial f(x) \iff (\zeta,0) \in \partial\varphi(x,Tx).$$

Because of the existence of x_0, we may apply Theorem 4.10 to φ, for any ζ as above. There results $(\alpha,\beta) \in N_{\operatorname{gr} T}(x,Tx)$ and $\gamma \in \partial g(Tx)$ such that

$$(\zeta,0) = (0,\gamma) + (\alpha,\beta).$$

The normal vector (α,β) satisfies

$$\langle \alpha, u-x\rangle + \langle \beta, Tu-Tx\rangle \leqslant 0 \;\; \forall u \in X,$$

which implies $\alpha = -T^*\beta$. We deduce $\zeta = T^*\gamma \in T^*\partial g(Tx)$, as required. □

Subdifferentials in Euclidean space. The basic theory of the subdifferential takes on a simpler form when we restrict attention to $\mathbb{R}^n$. We focus on this case in the remainder of this section.

4.14 Proposition. *Let $f : \mathbb{R}^n \to \mathbb{R}$ be convex. Then*

(a) *The graph of ∂f is closed:* $\zeta_i \in \partial f(x_i)$, $\zeta_i \to \zeta$, $x_i \to x \Longrightarrow \zeta \in \partial f(x)$;

(b) *For any compact subset S of $\mathbb{R}^n$, there exists M such that*

$$|\zeta| \leqslant M \;\; \forall \zeta \in \partial f(x),\; x \in S;$$

(c) *For any x, for any $\varepsilon > 0$, there exists $\delta > 0$ such that*

$$|y-x| < \delta \implies \partial f(y) \subset \partial f(x) + \varepsilon B.$$

Proof. Consider the situation described in (a). For any $y \in \mathbb{R}^n$, we have

$$f(y) - f(x_i) \geqslant \langle \zeta_i, y - x_i\rangle \;\; \forall i.$$

Taking limits, and bearing in mind that f is continuous (Cor. 2.35), we obtain in the limit $f(y) - f(x) \geqslant \langle \zeta, y-x\rangle$; thus, $\zeta \in \partial f(x)$.

We know that f is locally Lipschitz. An elementary argument using compactness shows that f is Lipschitz on bounded subsets of $\mathbb{R}^n$ (see Exer. 2.32). This, together with Prop. 4.6, implies (b). Part (c) follows from an argument by contradiction, using parts (a) and (b); we entrust this step to the reader. □

4.15 Exercise. (Mean value theorem) Let $f : \mathbb{R}^n \to \mathbb{R}$ be a convex function.

(a) We fix $x, v \in \mathbb{R}^n$ and we set $g(t) = f(x+tv)$ for $t \in \mathbb{R}$. Show that g is convex, and that

$$\partial g(t) = \big\langle \partial f(x+tv), v \big\rangle = \big\{ \langle \xi, v \rangle : \xi \in \partial f(x+tv) \big\}.$$

(b) Prove the following vaguely familiar-looking theorem: for all $x, y \in \mathbb{R}^n$, $x \neq y$, there exists $z \in (x, y)$ such that

$$f(y) - f(x) \in \big\langle \partial f(z), y - x \big\rangle.$$

(c) Let U be an open convex subset of $\mathbb{R}^n$. Use the above to prove the following subdifferential characterization of the Lipschitz property:

$$f \text{ is Lipschitz of rank } K \text{ on } U \iff |\zeta| \leqslant K \ \ \forall \zeta \in \partial f(x), \ \forall x \in U. \qquad \square$$

4.16 Proposition. *Let $f : \mathbb{R}^n \to \mathbb{R}$ be convex. Then f is differentiable at x if and only if $\partial f(x)$ is a singleton, and f is continuously differentiable in an open subset U if and only if $\partial f(x)$ reduces to a singleton for every $x \in U$.*

Proof. If f is differentiable at x, then Cor. 4.4 asserts that $\partial f(x)$ is the singleton $\{f'(x)\}$. Conversely, suppose that $\partial f(x)$ is a singleton $\{\zeta\}$. We proceed to prove that $f'(x) = \zeta$. Let x_i be any sequence converging to x ($x_i \neq x$). Then, by Exer. 4.15, there exist $z_i \in (x_i, x)$ and $\zeta_i \in \partial f(z_i)$ such that

$$f(x_i) - f(x) = \langle \zeta_i, x_i - x \rangle.$$

By part (c) of Prop. 4.14, the sequence ζ_i necessarily converges to ζ. Thus we have

$$\frac{|f(x_i) - f(x) - \langle \zeta, x_i - x \rangle|}{|x_i - x|} = \frac{|\langle \zeta_i - \zeta, x_i - x \rangle|}{|x_i - x|} \to 0,$$

whence $f'(x)$ exists and equals ζ. The final assertion is now easily verified with the help of part (c) of Prop. 4.14. $\square$

Strict convexity. Let U be a convex subset of X, and let $f : U \to \mathbb{R}$ be given. We say that f is *strictly convex* if the defining inequality of convexity is strict whenever it has any chance of being so:

$$x, y \in U, \ x \neq y, \ t \in (0,1) \implies f((1-t)x + ty) < (1-t)f(x) + t f(y).$$

It is easy to see that a C^2 function f of a single variable that satisfies $f''(t) > 0 \ \ \forall t$ is strictly convex (see Exer. 8.26 for an extension of this criterion to several dimensions). The role of strict convexity in optimization is partly to assure uniqueness of a minimum: the reader may check that a strictly convex function cannot attain its minimum at two different points.

4.17 Exercise. Let U be an open convex subset of $\mathbb{R}^n$, and let $f : U \to \mathbb{R}$ be convex.

(a) Prove that f is strictly convex if and only if

$$x, y \in U,\ x \neq y,\ \zeta \in \partial f(x) \implies f(y) - f(x) > \langle \zeta, y - x \rangle.$$

(b) Prove that f is strictly convex if and only if ∂f is injective, in the following sense:

$$x, y \in U,\ \partial f(x) \cap \partial f(y) \neq \emptyset \implies x = y. \qquad \square$$

4.2 Conjugate functions

Let X continue to designate a normed space, and let $f : X \to \mathbb{R}_\infty$ be a proper function (that is, $\operatorname{dom} f \neq \emptyset$). The *conjugate function* $f^* : X^* \to \mathbb{R}_\infty$ of f is defined by

$$f^*(\zeta) = \sup_{x \in X} \langle \zeta, x \rangle - f(x).$$

(One also refers to f^* as the *Fenchel* conjugate.) Note that the properness of f rules out the possibility that f^* has the value $-\infty$; we say then that f^* is well defined.

If $g : X^* \to \mathbb{R}_\infty$ is a proper function, its conjugate $g^* : X \to \mathbb{R}_\infty$ is defined by

$$g^*(x) = \sup_{\zeta \in X^*} \langle \zeta, x \rangle - g(\zeta).$$

Note that g^* is defined on X, and not on the dual of X^*, which we wish to avoid here. A special case arises when we take g to be f^*; then we obtain the *biconjugate* of f, namely the function $f^{**} : X \to \mathbb{R}_\infty$ defined as follows (when f^* is proper):

$$f^{**}(x) = \sup_{\zeta \in X^*} \langle \zeta, x \rangle - f^*(\zeta).$$

Since taking upper envelopes preserves both convexity and lower semicontinuity, it follows from the definition that f^* is convex lsc on the normed space X^*, and that f^{**} is convex lsc on X. The reader will observe that $f \leqslant g \implies f^* \geqslant g^*$.

4.18 Exercise.

(a) Show that for any function $f : X \to \mathbb{R}_\infty$, we have $f^{**} \leqslant f$.

(b) If f is proper, prove *Fenchel's inequality*:

$$f(x) + f^*(\zeta) \geqslant \langle \zeta, x \rangle \ \forall x \in X \ \forall \zeta \in X^*,$$

with equality if and only if $\zeta \in \partial f(x)$.

(c) Let f be the function $|x|^p/p$ on $\mathbb{R}^n$ $(1 < p < \infty)$. Calculate f^*, and show that Fenchel's inequality reduces in this case to *Young's inequality*:

$$u \bullet v \leqslant \frac{1}{p}|u|^p + \frac{1}{p_*}|v|^{p_*}, \quad u,v \in \mathbb{R}^n.$$

When does equality hold? □

4.19 Proposition. *Let $f : X \to \mathbb{R}_\infty$ be proper, and $c \in \mathbb{R}$ and $\zeta \in X^*$ be given. Then*

$$f(x) \geqslant \langle \zeta, x \rangle - c \;\; \forall x \in X \iff f^*(\zeta) \leqslant c.$$

If f is bounded below by a continuous affine function, then f^* and f^{**} are proper.

Proof. The first assertion, an equivalence, follows directly from the definition of f^*; it evidently implies that f^* is proper whenever f is bounded below by (majorizes, some would say) a continuous affine function. Since $f^{**} \leqslant f$, and since f is proper, we also deduce that f^{**} is proper. □

4.20 Proposition. *Let f be convex and lsc. Then f is bounded below by a continuous affine function. More explicitly, let x_0 be any point in X. If $x_0 \in \operatorname{dom} f$, then for any $\varepsilon > 0$, there exists $\zeta \in \operatorname{dom} f^*$ such that*

$$f(x) > f(x_0) + \langle \zeta, x - x_0 \rangle - \varepsilon \;\; \forall x \in X. \tag{1}$$

If $f(x_0) = +\infty$, then for any $M \in \mathbb{R}$, there exists $\zeta \in \operatorname{dom} f^*$ such that

$$f(x) > M + \langle \zeta, x - x_0 \rangle \;\; \forall x \in X. \tag{2}$$

Proof. Consider first the case $x_0 \in \operatorname{dom} f$. We apply the separation theorem to the point $(x_0, f(x_0) - \varepsilon)$ (a compact set) and the (closed) set $\operatorname{epi} f$. There results $\zeta \in X^*$ and $\lambda \in \mathbb{R}$ such that

$$\lambda r + \langle \zeta, x \rangle < \lambda f(x_0) - \lambda \varepsilon + \langle \zeta, x_0 \rangle \;\; \forall (x,r) \in \operatorname{epi} f.$$

It follows that $\lambda < 0$; we normalize to take $\lambda = -1$. This yields

$$f(x) > \langle \zeta, x \rangle + f(x_0) - \varepsilon - \langle \zeta, x_0 \rangle \;\; \forall x \in X,$$

the required inequality (which implies $\zeta \in \operatorname{dom} f^*$). The case $f(x_0) = +\infty$ is handled similarly, by separating the point (x_0, M) from $\operatorname{epi} f$. □

Notation. We denote by $\Gamma(X)$ the collection of all functions $f : X \to \mathbb{R}_\infty$ that are convex, lower semicontinuous, and proper. (This is classical notation in convex analysis.)

It follows from the two propositions above that when $f \in \Gamma(X)$, then f^* and f^{**} are both proper, which is a propitious context for studying conjugacy.

4.21 Theorem. (Moreau) *Let $f : X \to \mathbb{R}_\infty$ be a proper function. Then*

$$f \in \Gamma(X) \iff f^* \text{ is proper and } f = f^{**}.$$

Proof. When f^* is proper, then f^{**} is well defined (not $-\infty$), and it is convex and lsc as a consequence of the way it is constructed. Thus, if in addition we have $f = f^{**}$, then f belongs to $\Gamma\big(X\big)$.

Now for the converse; let $f \in \Gamma\big(X\big)$. Then f^* is proper by Prop. 4.19, and f^{**} is well defined. Since $f^{**} \leqslant f$, it suffices to establish that, for any given $x_0 \in X$, we have $f(x_0) \leqslant f^{**}(x_0)$. We reason by the absurd, supposing therefore that $f(x_0)$ is strictly greater than $f^{**}(x_0)$.

Let $M \in \mathbb{R}$ and $\varepsilon > 0$ satisfy $f(x_0) > M > f^{**}(x_0) + 2\varepsilon$. Any $\zeta \in \operatorname{dom} f^*$ admits $x_\zeta \in \operatorname{dom} f$ such that

$$f^*(\zeta) \leqslant \langle \zeta, x_\zeta \rangle - f(x_\zeta) + \varepsilon,$$

whence

$$f^{**}(x_0) \geqslant \langle x_0, \zeta \rangle - f^*(\zeta) \geqslant \langle \zeta, x_0 - x_\zeta \rangle + f(x_\zeta) - \varepsilon.$$

These facts lead to

$$f(x_0) > M > \langle \zeta, x_0 - x_\zeta \rangle + f(x_\zeta) + \varepsilon. \tag{3}$$

Now consider the case $f(x_0) < +\infty$. Then we choose $\zeta \in \operatorname{dom} f^*$ so that (1) holds. Then, using (3), we deduce

$$\begin{aligned} f(x_\zeta) &> f(x_0) + \langle \zeta, x_\zeta - x_0 \rangle - \varepsilon \\ &> \langle \zeta, x_0 - x_\zeta \rangle + f(x_\zeta) + \varepsilon + \langle \zeta, x_\zeta - x_0 \rangle - \varepsilon = f(x_\zeta), \end{aligned}$$

a contradiction. In the other case, when $f(x_0) = +\infty$, choose ζ so that (2) holds. Then (3) yields

$$M > \langle \zeta, x_0 - x_\zeta \rangle + f(x_\zeta) > \langle \zeta, x_0 - x_\zeta \rangle + M + \langle \zeta, x_\zeta - x_0 \rangle = M,$$

a contradiction which completes the proof. □

4.22 Corollary. *Let $g : X \to \mathbb{R}_\infty$ be a proper function which is bounded below by a continuous affine function. Then g^{**}, which is well defined, is the largest lsc convex function on X which is bounded above by g.*

Proof. We know that g^* is proper and that g^{**} is well defined, by Prop. 4.19. Since $g^{**} \leqslant g$, it is clear that g^{**} is indeed a convex lsc function bounded above by g. Let f be any other such function. Then $f \in \Gamma(X)$ and, by the theorem,

$$f \leqslant g \Longrightarrow f^* \geqslant g^* \Longrightarrow f = f^{**} \leqslant g^{**}.$$

□

In the corollary above, one can show that $\operatorname{epi} g^{**} = \overline{\operatorname{co}} \operatorname{epi} g$, which explains why g^{**} is sometimes referred to as the closed convex hull of g.

4.23 Corollary. *A function $g : X \to \mathbb{R}_\infty$ is convex and lsc if and only if there exists a family $\{\varphi_\alpha\}_\alpha$ of continuous affine functions on X whose upper envelope is g:*

$$g(x) = \sup_{\alpha} \varphi_\alpha(x) \ \ \forall x \in X.$$

Proof. An envelope of the indicated type is always convex and lsc, so if g has that form, it shares these properties. For the converse, we may suppose that g is proper (in addition to being convex and lsc). According to the theorem, we then have

$$g(x) = g^{**}(x) = \sup_{\zeta \in X^*} \langle \zeta, x \rangle - g^*(\zeta),$$

which expresses g as an upper envelope of the desired type. □

4.24 Exercise. We study conjugacy as it applies to indicators and support functions.

(a) Let S be a nonempty subset of X. Prove that $I_S{}^* = H_S$. Deduce from this that if S is closed and convex, then $H_S{}^* = I_S$.

(b) Let $\Sigma \subset X^*$ be nonempty, convex, and weak* closed. Prove that $H_\Sigma{}^* = I_\Sigma$. □

The following result allows us to recognize support functions.

4.25 Theorem. *Let $g : X \to \mathbb{R}_\infty$ be lsc, subadditive, and positively homogeneous, with $g(0) = 0$. Then there exists a unique nonempty weak* closed convex set Σ in X^* such that $g = H_\Sigma$. The set Σ is characterized by*

$$\Sigma = \big\{ \zeta \in X^* : g(v) \geqslant \langle \zeta, v \rangle \ \ \forall v \in X \big\},$$

and is weak* compact if and only if the function g is bounded on the unit ball.

Proof. Observe that the set Σ defined in the theorem statement is convex and weak* closed. We have (by definition)

$$g^*(\zeta) = \sup_{v \in X} \langle \zeta, v \rangle - g(v).$$

It follows from this formula and the positive homogeneity of g that $g^*(\zeta) = \infty$ if $\zeta \notin \Sigma$, and otherwise $g^*(\zeta) = 0$; that is, we have $g^* = I_\Sigma$. But $g \in \Gamma(X)$, which allows us to write (by Theorem 4.21)

$$g(x) = g^{**}(x) = \sup_{\zeta \in X^*} \langle \zeta, x \rangle - I_\Sigma(\zeta) = H_\Sigma(x).$$

The uniqueness of Σ follows from Cor. 3.13. If g is bounded on B, then it is clear that Σ is bounded, from its very definition. Then, as a weak* closed subset of some

ball, Σ is weak* compact (by Cor. 3.15). If, conversely, Σ is bounded, then $g = H_\Sigma$ is evidently bounded on the unit ball B. □

4.26 Corollary. *Let $f : X \to \mathbb{R}_\infty$ be a convex function which is Lipschitz near a point x. Then $f'(x;\cdot)$ is the support function of $\partial f(x)$:*

$$f'(x;v) = \max_{\zeta \in \partial f(x)} \langle \zeta, v \rangle \;\; \forall v \in X,$$

and f is Gâteaux differentiable at x if and only if $\partial f(x)$ is a singleton.

Proof. Consider the function $g(v) = f'(x;v)$. We invoke the convexity of f to write

$$f\big(x + \lambda[(1-t)v + tw]\big) \leqslant (1-t)f(x+\lambda v) + t f(x+\lambda w),$$

from which we deduce that $g((1-t)v+tw) \leqslant (1-t)g(v)+tg(w)$. Thus g is convex, and we see without difficulty that g is positively homogeneous and (from the Lipschitz condition) satisfies $|g(v)| \leqslant K\|v\| \;\; \forall v$. It follows that g is subadditive and continuous. Thus, g satisfies the hypotheses of Theorem 4.25. In light of Prop. 4.3, this implies the stated relation between $f'(x;\cdot)$ and $\partial f(x)$.

The last assertion of the corollary is now apparent, in view of Cor. 4.4. □

We remark that in more general circumstances than the above, the reduction of $\partial f(x)$ to a singleton does *not* imply Gâteaux differentiability; see Exer. 8.21.

4.27 Exercise. (Subdifferential inversion) Let $f \in \Gamma(X)$. For $\zeta \in \operatorname{dom} f^*$, the subdifferential $\partial f^*(\zeta)$ consists, by definition, of the points $x \in X$ such that

$$f^*(\xi) - f^*(\zeta) \geqslant \langle \xi - \zeta, x \rangle \;\; \forall \xi \in X^*.$$

Prove that

$$\zeta \in \partial f(x) \iff f(x) + f^*(\zeta) = \langle \zeta, x \rangle \iff x \in \partial f^*(\zeta). \qquad \square$$

4.28 Exercise. Prove that the points at which a function $f \in \Gamma(X)$ attains its minimum are those in $\partial f^*(0)$. □

4.3 Polarity

The geometric counterpart of conjugacy is *polarity*, an operation that plays an important role in the study of tangents and normals. That happens to be the context in which the reader has already made its acquaintance, in §1.4, in connection with defining the normal cone.

Let A be a subset of a normed space X. The **polar cone** of A (or, more simply, the polar), denoted $A^{\triangle}$, is defined by

$$A^{\triangle} = \{ \zeta \in X^* : \langle \zeta, x \rangle \leqslant 0 \ \forall x \in A \}.$$

It follows from the definition that $A^{\triangle}$ is a weak* closed convex cone. In the reverse direction, the polar of a subset Σ of X^* is defined by

$$\Sigma^{\triangle} = \{ x \in X : \langle \sigma, x \rangle \leqslant 0 \ \forall \sigma \in \Sigma \}.$$

The reader is asked to verify that we always have $A^{\triangle\triangle} \supset A$.

4.29 Exercise. Let A and Σ be nonempty cones in X and X^* respectively. Prove that $(I_A)^* = I_{A^{\triangle}}$ and $(I_{\Sigma})^* = I_{\Sigma^{\triangle}}$. □

4.30 Proposition. *Let A be a nonempty subset of X. Then A is a closed convex cone if and only if $A^{\triangle\triangle} = A$.*

Proof. The set $A^{\triangle\triangle}$ is always a closed convex cone, by construction. Thus, if it equals A, then A has these properties. Conversely, let A be a closed convex cone. We have $I_A = (I_A)^{**}$ by Theorem 4.21. However, by Exer. 4.29, we also have

$$(I_A)^{**} = (I_{A^{\triangle}})^* = I_{A^{\triangle\triangle}}.$$

Thus, $I_A = I_{A^{\triangle\triangle}}$, whence $A^{\triangle\triangle} = A$. □

4.31 Corollary. *Let A be a nonempty subset of X. Then the bipolar $A^{\triangle\triangle}$ of A is the smallest closed convex cone containing A.*

Proof. If K is a closed convex cone such that $K \supset A$, then, by the proposition, we deduce $K = K^{\triangle\triangle} \supset A^{\triangle\triangle}$. □

4.32 Corollary. *Let $x \in S$, where S is a convex subset of a normed space X. Then the tangent and normal cones at x are mutually polar*:

$$T_S(x) = N_S(x)^{\triangle}, \quad N_S(x) = T_S(x)^{\triangle}.$$

Proof. The second formula holds by definition. The first one then follows from Prop. 4.30, since $T_S(x)$ is (closed and) convex when S is convex (Prop. 2.9). □

We proceed now to focus on cones in X^*.

4.33 Exercise. Let Σ be a nonempty cone in X^*. Prove that $H_{\Sigma} = I_{\Sigma^{\triangle}}$. □

For bipolarity relative to X^*, the weak* topology plays a role once again, in allowing us to refer back to X:

4.34 Proposition. *Let Σ be a nonempty subset of X^*. Then Σ is a weak*closed convex cone if and only if $\Sigma^{\triangle\triangle} = \Sigma$.*

Proof. If $\Sigma^{\triangle\triangle} = \Sigma$, then it follows that Σ is a weak*closed convex cone, since $\Sigma^{\triangle\triangle}$ always has these properties. Let us now suppose that Σ is a weak*closed convex cone; we proceed to prove that it coincides with its bipolar. We know that $\Sigma \subset \Sigma^{\triangle\triangle}$. For purposes of obtaining a contradiction, assume that the opposite inclusion fails, and let $\zeta \in \Sigma^{\triangle\triangle} \setminus \Sigma$. By Theorem 3.2, there exists $x \in X$ such that

$$\langle \zeta, x \rangle > \langle \sigma, x \rangle \;\; \forall \sigma \in \Sigma.$$

(This is where the weak*closedness of Σ is used.) Since Σ is a cone, we get

$$\langle \zeta, x \rangle > 0 \geqslant \langle \sigma, x \rangle \;\; \forall \sigma \in \Sigma.$$

It follows that $x \in \Sigma^{\triangle}$ and thus $\zeta \notin \Sigma^{\triangle\triangle}$, the required contradiction. □

4.35 Exercise. Find $A^{\triangle}$ and $A^{\triangle\triangle}$, when $A \subset \mathbb{R}^2$ consists of the two points $(1,0)$ and $(1,1)$. □

4.4 The minimax theorem

Given a function $f(u,v)$ of two variables, its infimum can be calculated either jointly or successively, in either order:

$$\inf_u \inf_v f(u,v) = \inf_v \inf_u f(u,v) = \inf_{u,v} f(u,v).$$

When a supremum with respect to one of the variables is involved, however, the inf sup and the sup inf will differ in general. The following theorem (said to be of *minimax* type) gives conditions under which equality does hold. The first of many such results, due to von Neumann (see Exer. 4.37 below), figured prominently in game theory; they have since become a useful tool in analysis.

4.36 Theorem. (Ky Fan) *Let U and V be nonempty convex sets in (possibly different) vector spaces. Let $f\colon U \times V \to \mathbb{R}$ be a function such that*

$$u \mapsto f(u,v) \text{ is convex on } U \;\; \forall v \in V, \text{ and } v \mapsto f(u,v) \text{ is concave on } V \;\; \forall u \in U.$$

Suppose in addition that there is a topology on U for which U is compact and $f(\cdot,v)$ is lsc for each $v \in V$. Then we have

$$\sup_{v \in V} \min_{u \in U} f(u,v) = \min_{u \in U} \sup_{v \in V} f(u,v), \text{ where the case } \infty = \infty \text{ is admitted.}$$

Proof. We set $\alpha =$ the sup min, $\beta =$ the min sup. Since taking upper envelopes preserves lower semicontinuity, the hypotheses authorize the use of "min" here. It is easy to see that the inequality $-\infty < \alpha \leqslant \beta$ always holds; we may therefore restrict attention to the case $\alpha < \infty$. We now suppose that $\alpha < \beta$, and we derive a contradiction.

We claim that the sets $U(v) = \{u \in U : f(u,v) \leqslant \alpha\}$ satisfy $\bigcap_{v \in V} U(v) = \emptyset$. If this were not the case, then there would be a point $\bar{u}$ common to them all, so that $f(\bar{u},v) \leqslant \alpha \ \forall\, v \in V$. But then $\beta \leqslant \sup_v f(\bar{u},v) \leqslant \alpha < \beta$, absurd. Since the sets $U(v)$ are closed subsets of the compact space U, the finite intersection property must fail. Thus, there exist $v_1, \dots, v_n \in V$ such that $\bigcap_1^n U(v_i) = \emptyset$. Consider now the set

$$E = \big\{ x \in \mathbb{R}^n : \exists\, u \in U,\ r_i \geqslant 0 \text{ such that } x_i = f(u,v_i) + r_i,\ i = 1,2\dots,n \big\}.$$

It is easy to see that E is convex. Using the compactness of U and the lower semicontinuity of each function $f(\cdot,v_i)$, it is an exercise to show that the complement of E is open (that is, E is closed). We claim that E does not contain the point $p := (\alpha, \alpha, \dots, \alpha)$. For if it did, there would exist $u \in U$ and $r_i \geqslant 0$ such that

$$\alpha = f(u,v_i) + r_i \ \ \forall\, i.$$

But then $u \in \bigcap_1^n U(v_i)$, a contradiction. This proves the claim, and allows us to invoke Theorem 2.37 to separate $\{p\}$ and E. There results a vector $\zeta \in \mathbb{R}^n$ and a scalar γ such that

$$\zeta \bullet p < \gamma < \sum_{i=1}^n \zeta_i \big(f(u,v_i) + r_i\big) \ \ \forall\, u \in U,\ r_i \geqslant 0.$$

It follows that ζ is nonzero and has nonnegative components. We may normalize to arrange $\sum_1^n \zeta_i = 1$. Then the point $\bar{v} = \sum_1^n \zeta_i v_i$ belongs to V, and the previous inequality, combined with the concavity of f with respect to v, implies

$$\alpha = \zeta \bullet p < \min_{u \in U} f(u,\bar{v}) \leqslant \alpha.$$

This contradiction completes the proof. □

4.37 Exercise. Let M be an $m \times n$ matrix, and let S, T be closed, convex, nonempty subsets of $\mathbb{R}^m$ and $\mathbb{R}^n$ respectively, at least one of which is bounded. Then

$$\inf_{x \in S} \sup_{y \in T} \langle x, My \rangle = \sup_{y \in T} \inf_{x \in S} \langle x, My \rangle.$$ □

4.38 Exercise. Let X be a normed space, and let $g : X \to \mathbb{R}_\infty$ be a convex function, $x_0 \in X$, and $k > 0$. Then

$$\inf_{x \in X} \max_{\zeta \in kB_*} g(x) + \langle \zeta, x - x_0 \rangle = \max_{\zeta \in kB_*} \inf_{x \in X} g(x) + \langle \zeta, x - x_0 \rangle.$$ □

Chapter 5
Banach spaces

A normed space X is said to be a **Banach space** if its metric topology is *complete*. This means that every Cauchy sequence x_i in X, that is, one that satisfies

$$\lim_{i,j\to\infty} \|x_i - x_j\| = 0,$$

admits a limit in X: there exists a point $x \in X$ such that $\|x_i - x\| \to 0$.

Informally, the reader may understand the absence of such a point x as meaning that the space has a hole where x should be. For purposes of minimization, one of our principal themes, it is clear that the existence of minimizers is imperiled by such voids. Consider, for example, the vector space $\mathbb{Q}$ of rational numbers. The minimization of the function $(x^2-2)^2$ does *not* admit a solution over $\mathbb{Q}$, as the Greek mathematicians of antiquity were able to prove.

The existence of solutions to minimization problems is not the only compelling reason to require the completeness of a normed space, as we shall see. The property is essential in making available to us certain basic tools, such as uniform boundedness and weak compactness.

5.1 Completeness of normed spaces

It is easy to see that the completeness of a normed space is invariant with respect to equivalent norms.[1] Two other relevant properties that follow easily are the following: a closed subspace of a Banach space is itself a Banach space, and the Cartesian product of two Banach spaces is a Banach space. A less evident fact, one that we prove now, is that the dual space is always a Banach space.

[1] This in contrast to the purely metric case, in which completeness depends on the choice of metric, even among those inducing the same topology.

F. Clarke, *Functional Analysis, Calculus of Variations and Optimal Control*,
Graduate Texts in Mathematics 264, DOI 10.1007/978-1-4471-4820-3_5,

5.1 Theorem. *The dual X^* of a normed space X is a Banach space.*

Proof. It is understood, of course, that X^* is equipped with its usual dual norm $\|\cdot\|_*$. The issue here is to show that a given Cauchy sequence ζ_n in X^* admits a limit $\zeta \in X^*$. For each $x \in X$, it follows that $\langle \zeta_n, x \rangle$ is a Cauchy sequence in $\mathbb{R}$, since

$$\big|\langle \zeta_n, x \rangle - \langle \zeta_m, x \rangle\big| \leqslant \|\zeta_n - \zeta_m\|_* \|x\|.$$

Since $\mathbb{R}$ is complete, the sequence converges to a limit, denoted $\langle \zeta, x \rangle$. The function ζ so defined is a linear functional; we now show that it is continuous.

Fix N such that $m, n \geqslant N \Longrightarrow \|\zeta_m - \zeta_n\|_* \leqslant 1$. Then, for $n \geqslant N$, and for any $x \in B$, we have:

$$|\langle \zeta_n, x \rangle| \leqslant |\langle \zeta_n - \zeta_N, x \rangle| + |\langle \zeta_N, x \rangle| \leqslant 1 + \|\zeta_N\|_*.$$

Letting $n \to \infty$, we deduce $|\langle \zeta, x \rangle| \leqslant 1 + \|\zeta_N\|_*$, which implies that $\zeta \in X^*$.

To conclude, we must confirm that $\|\zeta_n - \zeta\|_* \to 0$. Let $\varepsilon > 0$. There exists N_ε such that $m, n \geqslant N_\varepsilon \Longrightarrow \|\zeta_n - \zeta_m\|_* < \varepsilon$. Fix $n \geqslant N_\varepsilon$. Then, for any $x \in B$ and $m > N_\varepsilon$, we have

$$|\langle \zeta_n - \zeta, x \rangle| \leqslant |\langle \zeta_n - \zeta_m, x \rangle| + |\langle \zeta_m - \zeta, x \rangle| \leqslant \varepsilon + |\langle \zeta_m - \zeta, x \rangle|.$$

Now $\langle \zeta_m - \zeta, x \rangle \to 0$ as $m \to \infty$, whence $|\langle \zeta_n - \zeta, x \rangle| \leqslant \varepsilon$. Since x is an arbitrary point in B, this reveals $\|\zeta_n - \zeta\|_* \leqslant \varepsilon \;\; \forall n \geqslant N_\varepsilon$. □

The proof of the theorem readily adapts to prove

5.2 Corollary. *If Y is a Banach space, then $L_C(X,Y)$ is a Banach space.*

We remark that any space that is isometric to a dual space is necessarily a Banach space, in view of the following fact:

5.3 Proposition. *Let $T : X \to Y$ be a norm-preserving map from a Banach space X to a normed space Y. Then T is closed: $T(S)$ is closed in Y whenever S is closed in X. If T is an isometry from X to Y, then Y is a Banach space.*

Proof. Let S be a closed subset of X; we prove that $T(S)$ is closed. Let x_i be a sequence in S such that Tx_i converges to a point $y \in Y$. To deduce that $T(S)$ is closed, we wish to prove that $y \in T(S)$.

Since Tx_i is a convergent sequence, it is Cauchy: $\|Tx_i - Tx_j\|_Y \to 0$ as $i, j \to \infty$. Because T is norm preserving, this implies that x_i is Cauchy. Since S is closed and X is complete, the sequence x_i admits a limit $u \in S$. By continuity, $Tx_i \to Tu = y$; we have proved, as required, that $y \in T(S)$. When T is an isometry (and thus surjective), the proof (applied to $S = X$) shows that Y is complete, and hence is a Banach space. □

5.4 Exercise. We return to the spaces of sequences defined in Example 1.6.

(a) Show that ℓ^1 is a Banach space.

(b) Use Theorem 5.1 and Prop. 5.3 to prove that ℓ^p $(1 < p \leqslant \infty)$ is a Banach space, and deduce that the spaces c and c_0 are Banach spaces.

(c) Show that ℓ^∞_c is not a Banach space. □

We now invite the reader to meet a chosen few of the many other Banach spaces that live in the world of analysis.

Spaces of continuous functions. We have met in Example 1.4 the normed space $C(K)$ of continuous functions on a compact metric space K. If f_i is a Cauchy sequence in $C(K)$, then, for each $x \in K$, the sequence $f_i(x)$ is Cauchy in $\mathbb{R}$. If we denote its limit by $f(x)$, then an argument not unlike the proof of Theorem 5.1 shows that $f \in C(K)$ and $\|f_i - f\|_{C(K)} \to 0$. Thus, $C(K)$ is a Banach space.

We observed earlier in Example 1.1 that when $K = [0,1]$, another norm on the vector space $C[0,1]$ is provided by

$$\|f\|_1 = \int_0^1 |f(t)|\,dt\,.$$

This norm is not complete, however:

5.5 Exercise. Show that the sequence $f_i(t) = [\min(2t,1)]^i$ is Cauchy relative to $\|\cdot\|_1$, but that f_i does not converge in that norm to an element of $C[0,1]$. □

The compactness of K is not used in an essential way to deduce that $C(K)$ is complete; it is there to guarantee that an element $f \in C(K)$ is bounded. An alternative to compactness is to incorporate this in the definition, as in the following:

5.6 Proposition. *Let X be a normed space and Y be a Banach space. Then the vector space $C_b(X,Y)$ of bounded continuous functions $g : X \to Y$ is a Banach space when equipped with the norm*

$$\|g\|_{C_b(X,Y)} = \sup_{x \in X} \|g(x)\|_Y\,.$$

(The function g is called "bounded" precisely when the right side above is finite.) Once again, the proof (which we omit) can be patterned after that of Theorem 5.1. We turn now to Lipschitz functions.

5.7 Proposition. *Let X be a normed space and Y be a Banach space. The vector space $Lip_b(X,Y)$ of bounded, Lipschitz functions $\varphi : X \to Y$ equipped with the norm*

$$\|\varphi\|_{Lip_b(X,Y)} = \|\varphi\|_{C_b(X,Y)} + \sup_{\substack{x,u \in X \\ x \neq u}} \frac{\|\varphi(x) - \varphi(u)\|_Y}{\|x-u\|_X}$$

is a Banach space.

Proof. Note the tacit assumption that X is not trivial. It is easy to see that we do have a normed space; we verify that it is complete. Let φ_n be a Cauchy sequence in $\mathrm{Lip}_b(X,Y)$. Then φ_n is also a Cauchy sequence in $C_b(X,Y)$, which is complete. In consequence, there exists $\varphi \in C_b(X,Y)$ such that $\|\varphi_n - \varphi\|_{C_b(X,Y)} \to 0$. Note that the functions φ_n have a common Lipschitz constant, which we denote by L. Passing to the limit in the inequality

$$\|\varphi_n(x) - \varphi_n(u)\|_Y \leqslant L\|x-u\|_X,$$

we see that φ is Lipschitz. Thus $\varphi \in \mathrm{Lip}_b(X,Y)$.

We complete the proof by showing that $\|\varphi_n - \varphi\|_{\mathrm{Lip}_b(X,Y)} \to 0$, which amounts to showing that

$$\lim_{n\to\infty} \sup_{x,u} \frac{\|(\varphi_n - \varphi)(x) - (\varphi_n - \varphi)(u)\|_Y}{\|x-u\|_X} = 0.$$

The left side of this desired equality coincides with

$$\begin{aligned}
&\lim_{n\to\infty} \sup_{x,u} \lim_{m\to\infty} \frac{\|(\varphi_n - \varphi_m)(x) - (\varphi_n - \varphi_m)(u)\|_Y}{\|x-u\|_X} \\
&\leqslant \lim_{n\to\infty} \lim_{m\to\infty} \sup_{x,u} \frac{\|(\varphi_n - \varphi_m)(x) - (\varphi_n - \varphi_m)(u)\|_Y}{\|x-u\|_X} \\
&\leqslant \lim_{n\to\infty} \lim_{m\to\infty} \|\varphi_n - \varphi_m\|_{\mathrm{Lip}_b(X,Y)} = 0,
\end{aligned}$$

since the sequence φ_n is Cauchy in $\mathrm{Lip}_b(X,Y)$. □

Lebesgue spaces and Sobolev spaces. It is shown in courses on integration that the Lebesgue space $L^p(\Omega)$ $(1 \leqslant p \leqslant \infty)$ is complete.[2] Since $\mathrm{AC}^p[a,b]$ (see Example 1.13) is isometric to $\mathbb{R} \times L^p(a,b)$, it follows that it too is a Banach space (by Prop. 5.3).

We introduce now certain functions that may be thought of as extensions of absolutely continuous functions to several dimensions. They play a central role in the multiple integral calculus of variations, and in partial differential equations. Let u be an element of $L^p(\Omega,\mathbb{R})$, where Ω is a nonempty open subset of $\mathbb{R}^n$. Then u is said to admit a **weak derivative** in L^p if there exists an element

[2] See for example Royden [36].

$$\big(g_1, g_2, \ldots, g_n\big) \in L^p(\Omega, \mathbb{R}^n)$$

such that, for each index i, we have

$$\int_\Omega u(x)\,\frac{\partial\varphi}{\partial x_i}(x)\,dx \;=\; -\int_\Omega g_i(x)\,\varphi(x)\,dx \quad \forall\,\varphi \in C_c^\infty(\Omega, \mathbb{R})\,.$$

Here, $C_c^\infty(\Omega, \mathbb{R})$ refers to the functions $\varphi : \Omega \to \mathbb{R}$ which admit partial derivatives of all orders, and which have compact support in Ω.

We shall prove in Chapter 6 that $C_c^\infty(\Omega, \mathbb{R})$ is a dense subset of $L^p(\Omega)$; this implies that the weak derivative $g = (g_1, g_2, \ldots, g_n)$ of u is unique when it exists. It is clear from the integration by parts formula that if u happens to be continuously differentiable, then its weak derivative is precisely its gradient. This justifies, by extension, the use of the notation Du for the weak derivative g of u. We also write $D_i u$ for the function g_i.

It is easy to see that if two functions u and v admit the weak derivatives Du and Dv, and if c, k are scalars, then $cu + kv$ admits the weak derivative $cDu + kDv$.

Let $1 \leqslant p \leqslant \infty$. The **Sobolev space** $W^{1,p}(\Omega)$ is by definition the vector space of all functions $u \in L^p(\Omega, \mathbb{R})$ which admit weak derivatives in L^p, equipped with the following norm:

$$\|u\|_{W^{1,p}(\Omega)} \;=\; \|u\|_{L^p(\Omega)} + \sum_{i=1}^n \|D_i u\|_{L^p(\Omega)}\,.$$

The space $W^{1,2}(\Omega)$ is usually denoted $H^1(\Omega)$.

5.8 Exercise. Let u_j be a sequence in $W^{1,p}(\Omega)$ $(1 \leqslant p < \infty)$ such that u_j converges weakly in $L^p(\Omega)$ to a limit u, and such that, for each $i = 1, 2, \ldots, n$, the sequence $D_i u_j$ converges weakly in $L^p(\Omega)$ to a limit v_i. Prove that $u \in W^{1,p}(\Omega)$, and that $Du = (v_1, v_2, \ldots, v_n)$. □

The following exercise shows that, when $n = 1$, with $\Omega = (a, b)$, the functions u in $W^{1,1}(\Omega)$ are essentially the elements of $\mathrm{AC}[a, b]$.

5.9 Exercise. Prove that u lies in $W^{1,1}(a, b)$ if and only if there is a function f in $\mathrm{AC}[a, b]$ such that $u(t) = f(t)$, $t \in [a, b]$ a.e., in which case we have $Du = f'$. □

In a context such as this, bearing in mind that the elements of $L^p(a, b)$ are really equivalence classes, it is common to say that u admits a continuous *representative*. When $n > 1$, it is no longer the case that an element of $W^{1,p}(\Omega)$ necessarily admits a continuous representative, or even a locally bounded one.

5.10 Exercise. Prove that $W^{1,p}(\Omega)$ is a Banach space. □

The uniform boundedness principle. The following very useful fact, also known as the Banach-Steinhaus theorem, plays a central role in the theory, and should be part of the reader's repertoire.

5.11 Theorem. *Let X be a Banach space and Y a normed space. Let Γ be a collection of operators in $L_C(X,Y)$. If Γ is simply bounded (meaning that, for every $x \in X$, the image set $\{\Lambda x : \Lambda \in \Gamma\}$ is bounded in Y), then Γ is uniformly bounded: there exists $M \geqslant 0$ such that*

$$\|\Lambda\|_{L_C(X,Y)} \leqslant M \;\; \forall \Lambda \in \Gamma.$$

The theorem affirms that if the set Γ in $L_C(X,Y)$ is simply bounded, then it is, quite simply, bounded.

Proof. For each integer n, we define a subset of X by

$$F_n = \big\{ x \in X : \|\Lambda x\|_Y \leqslant n \; \forall \Lambda \in \Gamma \big\}.$$

Note that F_n is closed, since, for each $\Lambda \in \Gamma$, the function $x \mapsto \|\Lambda x\|_Y$ is continuous. By hypothesis, we have $X = \cup_{n \geqslant 1} F_n$. Since X is complete, Baire's theorem[3] tells us that, for some N, there exists $z \in X$ and $r > 0$ such that $B(z,r) \subset F_N$. Thus we have

$$\|\Lambda(z+ru)\|_Y \leqslant N \;\; \forall u \in B,\; \Lambda \in \Gamma.$$

We deduce

$$r\|\Lambda u\|_Y \leqslant N + \|\Lambda z\|_Y \leqslant N + N = 2N \;\; \forall u \in B,\; \Lambda \in \Gamma.$$

Taking the supremum over $u \in B$, we obtain

$$\|\Lambda\| = \sup_{u \in B} \|\Lambda u\|_Y \leqslant (2N)/r =: M \;\; \forall \Lambda \in \Gamma.$$

□

The next three exercises are rather immediate corollaries of the uniform boundedness principle. In each case, one may find counterexamples to show the necessity of the completeness hypothesis.

5.12 Exercise. Let Λ_n be a sequence of continuous linear operators mapping a Banach space X to a normed space Y. Suppose that for each $x \in X$, the limit $\Lambda x := \lim_{n \to \infty} \Lambda_n x$ exists. Prove that the (linear) mapping Λ is continuous. □

5.13 Exercise. Prove that a weakly compact subset of a normed space is bounded. Deduce that a weakly convergent subsequence in a normed space is bounded. □

---

[3] We have in mind: **Theorem.** *Let (E,d) be a complete metric space, and F_n a sequence of closed subsets of E such that* $\operatorname{int}\{\bigcup_n F_n\} \neq \emptyset$. *Then there exists N such that* $\operatorname{int} F_N \neq \emptyset$.

5.14 Exercise. Let X be a Banach space and let ζ_n be a sequence in X^* converging weak* to 0 ; that is, such that for every $x \in X$, we have $\lim_{n\to\infty}\langle \zeta_n, x\rangle = 0$. Prove that ζ_n is bounded in X^*. □

Support functions and boundedness. As we now see, in a normed space, and in the dual of a Banach space, the boundedness of a (nonempty) set is equivalent to its support function being finite-valued.

5.15 Proposition. *Let X be a normed space. A subset S of X is bounded if and only if*

$$H_S(\zeta) := \sup_{x\in S} \langle \zeta, x\rangle < \infty \;\; \forall\, \zeta \in X^*.$$

If X is a Banach space, then a subset Σ of X^* is bounded if and only if

$$H_\Sigma(x) := \sup_{\sigma\in\Sigma} \langle \sigma, x\rangle < \infty \;\; \forall\, x \in X.$$

Proof. We may assume that the sets involved are nonempty. If S is bounded, then $S \subset rB$ for some r, whence

$$H_S(\zeta) \leqslant \sup_{x\in rB} \langle \zeta, x\rangle \leqslant r\|\zeta\|_*\|x\|.$$

Thus H_S is finite-valued; a similar argument is valid for H_Σ.

Suppose now that H_S is finite-valued, and let us deduce from this that S is bounded. Every $x \in S$ engenders an element Λ_x of $L_C(X^*, \mathbb{R})$ via $\Lambda_x(\zeta) = \langle \zeta, x\rangle$, and one has $\|\Lambda_x\| = \|x\|$. The family $\Gamma = \{\Lambda_x\}_{x\in S}$ is simply bounded since, for a given $\zeta \in X^*$, we have

$$\inf_{x\in S}\langle \zeta, x\rangle = -H_S(-\zeta) > -\infty, \quad \sup_{x\in S}\langle \zeta, x\rangle = H_S(\zeta) < +\infty.$$

Then, since X^* is a Banach space (Theorem 5.1), the family Γ is bounded by Theorem 5.11. Because we have $\|\Lambda_x\| = \|x\|$, it follows that S is bounded.

The argument for Σ is similar, using the family $\{\Lambda_\zeta\}$, where $\Lambda_\zeta(x) = \langle \zeta, x\rangle$. Note that now we need to posit that X is complete, however, in order to call upon the uniform boundedness principle. □

5.16 Exercise. In the second assertion of the preceding theorem, X is taken to be a Banach space. Show that this is essential: give an example of an unbounded set Σ in the dual of a normed space X whose support function H_Σ is everywhere finite-valued. □

Continuity of convex functions. In a Banach space, the convexity of a lower semicontinuous function automatically implies its Lipschitz continuity (compare with Theorem 2.34):

5.17 Theorem. *Let X be a Banach space, and let $f : X \to \mathbb{R}_\infty$ be convex and lsc. Then f is locally Lipschitz in the set* $\operatorname{int} \operatorname{dom} f$.

Proof. [Exercise] Let $x_0 \in \operatorname{int} \operatorname{dom} f$. In view of Theorem 2.34, it suffices to prove that the (closed) set

$$C = \big\{ y \in X : f(x_0 + y) \leqslant f(x_0) + 1 \big\}$$

has nonempty interior.

(a) Prove that, for every point $z \in X$, there exists $t > 0$ such that $tz \in C$.

(b) Invoke Baire's theorem to deduce that $\operatorname{int} C \neq \emptyset$. □

That the lower semicontinuity hypothesis in Theorem 5.17 is required follows from considering a discontinuous linear functional (which cannot be lsc).

5.18 Exercise. We set $X = \ell_c^\infty$ and, for $x = (x_1, x_2, \dots) \in X$, we define $f(x)$ to equal $\sum_i |x_i|$. Show that the function $f : X \to \mathbb{R}$ is convex and lsc, but fails to be continuous. □

We deduce from this example that the completeness hypothesis in Theorem 5.17 is also essential.

5.2 Perturbed minimization

It is clear that the completeness of a normed space is an important factor in the quest to identify a set of ingredients allowing us to affirm that minimization problems admit a solution. But completeness is not enough: we generally require a pinch of compactness as well. Later, we shall find a way to obtain it by exploiting the weak topology. Here, however, we examine a more modern consideration that has been very productive, an approach in which the original problem (that may not admit a minimum) is slightly perturbed to obtain a new problem that *does* have one. We refer to a theorem making this type of assertion as a *minimization principle*, for short; a more descriptive but even longer term would be "perturbed minimization principle."

In 1944, the analyst J. E. Littlewood formulated his famous "three principles" in analysis, that the reader may very well have seen. (One is that every measurable function is nearly continuous.[4]) We venture to propose a fourth principle: every function that is bounded below nearly attains a minimum. Of course, the trick is how to interpret the word "nearly" in making this precise.

[4] The other two are: every convergent sequence of functions is nearly uniformly convergent, and every measurable set is nearly a finite union of intervals.

We shall prove a first such result in this section, and several others in Chapter 7. The later ones will require that the underlying space be "smooth" in some sense. In contrast, the following minimization principle is valid in the general setting of a complete metric space.

5.19 Theorem. (Ekeland) *Let (E,d) be a complete metric space and let the function $f : E \to \mathbb{R}_\infty$ be proper, lsc, and bounded below. If $\varepsilon > 0$ and $u \in E$ satisfy $f(u) \leqslant \inf_E f + \varepsilon$, then, for any $\lambda > 0$, there exists $z \in E$ such that*

(a) $f(z) \leqslant f(u)$

(b) $d(u,z) \leqslant \lambda$

(c) $f(w) + (\varepsilon/\lambda)\, d(w,z) > f(z) \;\; \forall w \in E,\ w \neq z$.

We may summarize the conclusions of the theorem as follows: there is a point z which is λ-close to u (satisfies (b)), which is "at least as good as u" (satisfies (a)), such that the perturbed function $f(\cdot) + (\varepsilon/\lambda)\, d(\cdot, z)$ attains a (unique) minimum at z (conclusion (c)).

Proof. We fix $\alpha > 0$. The *partial order of Bishop-Phelps* on $E \times \mathbb{R}$ is defined as follows:

$$(x_1, r_1) \preccurlyeq (x_2, r_2) \iff r_1 + \alpha\, d(x_1, x_2) \leqslant r_2 .$$

The reader will easily verify that this relation is reflexive (we have $(x,r) \preccurlyeq (x,r)$) and transitive: if $(x,r) \preccurlyeq (y,s)$ and $(y,s) \preccurlyeq (z,t)$, then $(x,r) \preccurlyeq (z,t)$. It is also antisymmetric: if $(x,r) \preccurlyeq (y,s)$ and $(y,s) \preccurlyeq (x,r)$, then $(x,r) = (y,s)$.

We also remark that when $E \times \mathbb{R}$ is equipped with the product topology, the partial order $\preccurlyeq$ is sequentially closed, in the following sense:

$$\begin{aligned}
(x_i, r_i) \preccurlyeq (y,s),\ \ (x_i, r_i) \to (x,r) &\implies (x,r) \preccurlyeq (y,s) \\
(x,r) \preccurlyeq (y_i, s_i),\ \ (y_i, s_i) \to (y,s) &\implies (x,r) \preccurlyeq (y,s) .
\end{aligned}$$

Lemma. *Let P be a closed nonempty subset of $E \times \mathbb{R}$ which is bounded below, in the sense that*

$$\inf \bigl\{ r \in \mathbb{R} : \exists\, x \in X,\ (x,r) \in P \bigr\} > -\infty .$$

Then P contains a minimal element (x_*, r_*); that is, an element (x_*, r_*) having the property that the only point $(x,r) \in P$ which satisfies $(x,r) \preccurlyeq (x_*, r_*)$ is (x_*, r_*) itself.

Proof. We wish to invoke Zorn's lemma, in its "minimal element" rather than "maximal element" form. Accordingly, to verify that P is inductive, we show that any nonempty totally ordered subset Q of P admits a minorant in P.

Set $r_* = \inf \bigl\{ r : (x,r) \in Q \bigr\}$, a number in $\mathbb{R}$. Suppose first that there is a point of the form (x_*, r_*) in Q. Then it is clear that (x_*, r_*) is a minorant for Q:

$$(x_*, r_*) \preccurlyeq (x, r) \;\; \forall (x, r) \in Q,$$

since, for $(x, r) \in Q$ different from (x_*, r_*), we cannot have $(x, r) \preccurlyeq (x_*, r_*)$ (by the way r_* is defined), and since Q is totally ordered.

We may therefore limit attention to the second case, in which the level $r = r_*$ is not attained in Q. Then there is a sequence (x_i, r_i) in Q such that r_i strictly decreases to r_*. From the fact that Q is totally ordered, we deduce

$$r_{i+1} - r_i + \alpha\, d(x_i, x_{i+1}) \leqslant 0 \;\; \forall i \geqslant 1$$

(the opposite being impossible). It follows that x_i is a Cauchy sequence. Since E is complete, there exists $x_* \in E$ such that $x_i \to x_*$; we also have $(x_i, r_i) \to (x_*, r_*)$. We now claim that (x_*, r_*), which lies in P since P is closed, is the minorant of Q that we seek.

Let (x, r) be any point in Q. Suppose that $(x, r) \preccurlyeq (x_i, r_i)$ infinitely often. Passing to the limit, we derive $(x, r) \preccurlyeq (x_*, r_*)$; that is $r + \alpha\, d(x, x_*) \leqslant r_*$. This implies $r = r_*$, contradicting our assumption that r_* is not attained in Q. Thus we must have $(x_i, r_i) \preccurlyeq (x, r)$ for all i sufficiently large, and as a result, $(x_*, r_*) \preccurlyeq (x, r)$. This shows that (x_*, r_*) is a minorant for Q. The lemma is proved. □

We now prove the theorem by a direct application of the lemma, in which we take

$$P = \big\{ (z, r) \in \operatorname{epi} f \, : \, (z, r) \preccurlyeq \big(u, f(u)\big) \big\}.$$

Then P is nonempty, closed because f is lsc, , and bounded below in the sense of the lemma, since f is bounded below. We take $\alpha = \varepsilon/\lambda$ in the definition of $\preccurlyeq$.

It follows that P admits a minimal element (z, r), and the minimality implies that $r = f(z)$. Since $(z, f(z)) \in P$, we have $(z, f(z)) \preccurlyeq (u, f(u))$, that is

$$f(v) + (\varepsilon/\lambda)\, d(u, z) \leqslant f(u), \tag{$*$}$$

which implies conclusion (a) of the theorem.

Let w satisfy

$$w \in E\,, \;\; w \in \operatorname{dom} f, \;\; w \neq z.$$

If $(w, f(w)) \in P$, then we do *not* have $(w, f(w)) \preccurlyeq (z, f(z))$, since $(z, f(z))$ is minimal relative to P. Thus

$$f(w) + (\varepsilon/\lambda)\, d(w, z) > f(z),$$

which is the inequality in conclusion (c) of the theorem. In the opposite case, when $(w, f(w)) \notin P$, the relation $(w, f(w)) \preccurlyeq (u, f(u))$ necessarily fails (by definition of P), so that

$$\begin{aligned} f(w) &> f(u) - (\varepsilon/\lambda)\, d(w,u) \\ &\geqslant f(z) + (\varepsilon/\lambda)\, d(u,z) - (\varepsilon/\lambda)\, d(w,u) \text{ (by } (*)) \\ &\geqslant f(z) - (\varepsilon/\lambda)\, d(w,z)\,, \end{aligned}$$

by the triangle inequality. We obtain once again the inequality in (c), which holds therefore for all $w \neq v$.

There remains conclusion (b). From $f(u) \leqslant \inf_E f + \varepsilon$ we deduce $f(z) \geqslant f(u) - \varepsilon$. Together with $(*)$, this implies

$$f(z) \geqslant f(z) + (\varepsilon/\lambda)\, d(u,z) - \varepsilon\,,$$

whence $d(u,z) \leqslant \lambda$. □

Fermat's rule asserts that $f'(x)$ is zero at a minimizer x. It is *not* true that "$f'(x)$ is almost zero when x is almost a minimizer." However, as the following may be interpreted as saying, if x is almost minimizing, then there is a point which is almost the same point, at which the derivative is almost zero. (The result also estimates the three "almosts.")

5.20 Corollary. *Let $f: X \to \mathbb{R}$ be differentiable and bounded below, where X is a Banach space. Let $\varepsilon > 0$ and $x \in X$ satisfy $f(x) \leqslant \inf_X f + \varepsilon$. Then there exists x_ε such that*

$$\|x - x_\varepsilon\| \leqslant \sqrt{\varepsilon}\,, \quad f(x_\varepsilon) \leqslant f(x)\,, \quad \|f'(x_\varepsilon)\|_* \leqslant \sqrt{\varepsilon}\,.$$

Proof. We apply Theorem 5.19 with $u = x,\ E = X,\ \lambda = \sqrt{\varepsilon}$. We derive the existence of $x_\varepsilon \in B(x, \sqrt{\varepsilon})$ such that $f(x_\varepsilon) \leqslant f(x)$ and

$$f(y) + \sqrt{\varepsilon}\, \|y - x_\varepsilon\| \geqslant f(x_\varepsilon) \ \ \forall\, y \in X.$$

For fixed $w \in X$, let us substitute $y = x_\varepsilon + tw\ (t > 0)$ in this inequality; then we divide by t and let t decrease to 0. This leads to $\langle f'(x_\varepsilon), w\rangle \geqslant -\sqrt{\varepsilon}\, \|w\|$. Since w is arbitrary, we conclude that $\|f'(x_\varepsilon)\|_* \leqslant \sqrt{\varepsilon}$. □

Another consequence of the theorem bears upon the density of points at which the subdifferential of a convex function is nonempty.[5]

5.21 Proposition. *Let X be a Banach space, and let $f: X \to \mathbb{R}_\infty$ be convex and lsc. Then $\partial f(x) \neq \emptyset$ for all x in a dense subset of* $\operatorname{dom} f$. *More precisely, for every $x \in \operatorname{dom} f$ and $\delta > 0$, there exists x_δ satisfying*

$$\|x_\delta - x\| \leqslant \delta\,, \ |f(x) - f(x_\delta)| \leqslant \delta\,, \ \partial f(x_\delta) \neq \emptyset.$$

[5] Prop. 4.6 established the nonemptiness of the subdifferential at points of continuity.

Proof. We consider first the case in which f satisfies an extra assumption: f is bounded below.

Let $x \in \operatorname{dom} f$ be given. By the lower semicontinuity of f, there exists $\eta > 0$ such that $f \geqslant f(x) - \delta$ on $B(x,\eta)$. Now invoke Theorem 5.19 with $\varepsilon = f(x) - \inf_X f$ and $\lambda = \min(\delta,\eta)$. (We may assume $\varepsilon > 0$, for otherwise we have $0 \in \partial f(x)$, and there is nothing left to prove.)

There results a point $x_\delta \in B(x,\lambda) \subset B(x,\delta)$ which minimizes the function

$$u \mapsto f(u) + (\varepsilon/\lambda)\| u - x_\delta \|$$

over X, and which satisfies $f(x_\delta) \leqslant f(x)$. Note that $f(x_\delta) \geqslant f(x) - \delta$, whence $|f(x) - f(x_\delta)| \leqslant \delta$. By Theorem 4.10, we have

$$0 \in \partial f(x_\delta) + B(0, \varepsilon/\lambda),$$

so that $\partial f(x_\delta) \neq \emptyset$.

Consider now the general case. By Cor. 4.23, there exists $\zeta \in X^*$ such that the function $f(u) - \langle \zeta, u \rangle$ is bounded below. It is a simple exercise to apply the case proved above to this function, in order to obtain the required conclusion; we entrust the details to the reader. □

The decrease principle. Let $f : X \to \mathbb{R}$ be a differentiable function on a Banach space. If $f'(x) \neq 0$, then clearly we have, for any $r > 0$,

$$\inf_{B(x,r)} f < f(x).$$

This is simply Fermat's rule in contrapositive form. It is possible to give a *calibrated* version of this fact, as we now see. Note that when the function f is differentiable at x, there is equivalence between the derivative being nonzero and the existence of "directions of decrease," in the following sense:

$$\| f'(u) \|_* \geqslant \delta \iff \inf_{\|v\| \leqslant 1} f'(u;v) \leqslant -\delta \iff \inf_{\|v\|=1} f'(u;v) \leqslant -\delta.$$

The theorem below postulates pointwise decrease along these lines, but in a weaker sense, and without supposing that f is differentiable.

5.22 Theorem. (Decrease principle) *Let $f : X \to \mathbb{R}_\infty$ be lower semicontinuous. Suppose that for some $x \in \operatorname{dom} f$ and positive numbers δ and r we have*

$$\liminf_{w \to u,\, w \neq u} \frac{f(w) - f(u)}{\| w - u \|} \leqslant -\delta \qquad \forall\, u \in B^\circ(x,r) \cap \operatorname{dom} f.$$

Then

$$\inf \{ f(u) : u \in B^\circ(x,r) \} \leqslant f(x) - \delta r.$$

Proof. Assume that the conclusion of the theorem fails. Then there exists η in $(0, r/2)$ sufficiently small so that

$$\inf \{ f(u) : u \in B(x, r-\eta) \} + \delta(r-2\eta) \geqslant f(x).$$

We deduce from Theorem 5.19, with $E = B(x, r-\eta)$, that for any $\lambda \in (0, r-\eta)$ there exists $z \in B(x, \lambda)$ such that

$$f(w) + \frac{\delta(r-2\eta)}{\lambda} \| w - z \| > f(z) \quad \forall w \in B(x, r-\eta), w \neq z.$$

Since $\| z - x \| \leqslant \lambda < r - \eta$, we derive from the hypothesis of the theorem (invoked for $u = z$), and from the preceding inequality, the following:

$$-\delta \geqslant \liminf_{w \to z,\, w \neq z} \frac{f(w) - f(z)}{\| w - z \|} \geqslant \frac{\delta(2\eta - r)}{\lambda}.$$

If λ is taken greater than $r - 2\eta$, this provides the sought-for contradiction. □

5.23 Exercise. Let $f : X \to \mathbb{R}_\infty$ be lower semicontinuous and admit directional derivatives at points in $\operatorname{dom} f$. Suppose that for some $x \in \operatorname{dom} f$ and positive numbers δ and r we have

$$\inf_{\| v \| \leqslant 1} f'(u; v) \leqslant -\delta \;\; \forall u \in B^\circ(x, r) \cap \operatorname{dom} f.$$

Prove that the conclusion of Theorem 5.22 holds. □

5.3 Open mappings and surjectivity

The following result bears upon the surjectivity of an operator.

5.24 Theorem. *Let X be a Banach space and Y a normed space, and let $T : X \to Y$ be a continuous linear operator. Then*

$$\delta > 0, \;\operatorname{cl}(TB_X) \supset \delta B_Y \implies TB_X^\circ \supset \delta B_Y^\circ.$$

Proof.

Fix any $\alpha \in \delta B_Y^\circ$, $\alpha \neq 0$. We proceed to show that $\alpha \in TB_X^\circ$, which proves the theorem (since $0 \in TB_X^\circ$ is evident). Reasoning *ad absurdum*, let us suppose that $\alpha \notin TB_X^\circ$ and derive a contradiction.

Consider the convex function $f(x) = \| Tx - \alpha \|$. (The norm is that of Y.) With an eye to applying Theorem 5.22, we claim that

$$\inf_{\|v\| \leqslant 1} f'(u;v) \leqslant -\delta \;\; \forall u \in B_X^{\circ}.$$

Indeed, fix $u \in B_X^{\circ}$. From the hypothesis $\operatorname{cl}\big(TB_X\big) \supset \delta B_Y$, we know that, for any $\eta > 0$, there exist $v \in B_X$ and $z \in B_Y$ such that

$$Tv = \delta \frac{\alpha - Tu}{\|\alpha - Tu\|} + \eta z$$

(we have used the assumption $\alpha \notin TB_X^{\circ}$ in writing this). For $t > 0$ we have

$$T(u+tv) - \alpha = Tu - \alpha + tTv = (Tu - \alpha)\left(1 - \frac{t\delta}{\|\alpha - Tu\|}\right) + t\eta z$$

so that, for t small enough,

$$\|T(u+tv) - \alpha\| \leqslant \|Tu - \alpha\|\left(1 - \frac{t\delta}{\|\alpha - Tu\|}\right) + t\eta$$

and thus

$$\frac{f(u+tv) - f(u)}{t} = \frac{\|T(u+tv) - \alpha\| - \|Tu - \alpha\|}{t} \leqslant -\delta + \eta.$$

Proposition 2.22 and the arbitrariness of $\eta > 0$ then imply that $f'(u;v) \leqslant -\delta$, which establishes the claim.

It now follows from the decrease principle (see Exer. 5.23) that

$$0 \leqslant \inf\left\{\|Tu - \alpha\| : u \in B_X^{\circ}\right\} \leqslant f(0) - \delta = \|\alpha\| - \delta < 0,$$

which is the desired contradiction. □

A mapping $T : X \to Y$ is called *open* if $T(U)$ is an open set in Y whenever U is an open set in X.

5.25 Exercise. Let X and Y be normed spaces and $T : X \to Y$ a linear map. Show that T is open if and only if $T(B_X)$ contains a neighborhood of 0. Deduce from this that an open linear mapping is surjective. □

The converse of this last conclusion is (another) celebrated theorem of Banach.

5.26 Theorem. (Open mapping theorem) *Let X and Y be Banach spaces, and let $T : X \to Y$ be a continuous linear operator which is surjective. Then T is an open mapping. If in addition T is injective, then the inverse mapping T^{-1} is continuous.*

Proof. Define $C = \operatorname{cl} T(B_X)$. Since T is linear and surjective, we have

$$TX = Y = \bigcup_{i \geqslant 1} \operatorname{cl} T(iB_X) = \bigcup_{i \geqslant 1} iC.$$

Because Y is a Banach space, and hence complete, Baire's theorem tells us that there is a point $\alpha \in \operatorname{int} C$. But then C (which is symmetric and convex) is a neighborhood of 0, by the following argument:

$$0 \in \operatorname{int}\big(C/2-\alpha/2\big) \subset \operatorname{int}\big(C/2-C/2\big) = \operatorname{int}\big(C/2+C/2\big) = \operatorname{int} C.$$

By Theorem 5.24, $T(B_X)$ contains a neighborhood of 0. This, together with Exer. 5.25, yields the first part of the theorem.

The second part is immediate: for any open set U in X, $(T^{-1})^{-1}(U) = T(U)$ is open, whence the continuity of T^{-1}. □

5.27 Exercise. Let X be a Banach space relative to two norms $\|\cdot\|_1$ and $\|\cdot\|_2$. Suppose there exists $c > 0$ such that

$$\|x\|_1 \leqslant c\|x\|_2 \;\; \forall x \in X.$$

Prove that the norms are equivalent; that is, for some constant $d > 0$, we have $\|x\|_2 \leqslant d\|x\|_1 \;\; \forall x \in X$. (This fails without completeness; the norms of Example 1.1 provide a counter-example.) □

The **graph** of a mapping $T : X \to Y$ refers to the set

$$\operatorname{gr} T = \big\{(x, Tx) \in X \times Y : x \in X\big\}.$$

5.28 Theorem. (Closed graph theorem) *Let $T : X \to Y$ be a linear mapping between two Banach spaces X and Y. Then T is continuous if and only if its graph is closed.*

Proof. It is clear that the continuity of T implies that $G := \operatorname{gr} T$ is closed. For the converse, let G be closed. Then G is a closed subspace of the Banach space $X \times Y$, thus a Banach space itself.

The projection $\pi_X : G \to X$ defined by $\pi_X(x,y) = x$ is continuous and bijective. By the open mapping theorem, π_X^{-1} is continuous. But we may write $T = \pi_Y \circ \pi_X^{-1}$, where $\pi_Y : X \times Y \to Y$ is the projection on Y (also continuous). It follows that T, as a composition of continuous functions, is continuous. □

5.29 Exercise. Let X and Y be Banach spaces and $T : X \to Y$ a linear mapping such that

$$x_n \to 0, \; \Lambda \in Y^* \implies \Lambda T x_n \to 0.$$

Then T is continuous. □

5.4 Metric regularity

We consider now the solutions of a given equation

$$\varphi(x,y) = 0,$$

where $\varphi : X \times Y \to [0,\infty)$ is continuous and X, Y are Banach spaces. Our study focuses upon the stability of the set of solutions x to the equation for given values of the parameter y.

We are given a base solution $(\bar{x},\bar{y})$ of the underlying equation: a point such that $\varphi(\bar{x},\bar{y}) = 0$. In many cases, as we shall see, an important question is whether the equation still admits a solution when the parameter value $\bar{y}$ is perturbed; that is, changed to a nearby value y.

Notation: For each fixed y, the (possibly empty) set of solutions $x \in X$ of the equation $\varphi(x,y) = 0$ is denoted by $S(y)$.

Thus, we have $\bar{x} \in S(\bar{y})$, and the first question is whether $S(y)$ is nonempty when y is sufficiently near $\bar{y}$. More than this, however, we would like to be able to assert a *stability* property: namely, that when y is near $\bar{y}$, then $S(y)$ contains a point which is "proportionally close" to the original solution $\bar{x}$. We shall interpret this last phrase to mean that (for a suitable constant K) we have

$$d(\bar{x}, S(y)) \leqslant K\varphi(\bar{x}, y),$$

where $d(x,S)$ denotes the distance from x to the set S. (Recall that the function φ has nonnegative values.) Thus, it is φ itself which calibrates how close the perturbed solution should be. Note that such an inequality, when it holds, forces $S(y)$ to be nonempty, since $d(\bar{x}, S(y)) = +\infty$ when $S(y) = \emptyset$. In fact, we shall be able to assert stability with respect to perturbations of the x variable as well as the y variable; thus, *joint* stability with respect to (x,y). The conclusion will be of the form

$$d(x, S(y)) \leqslant K\varphi(x,y) \quad \text{for } (x,y) \text{ near } (\bar{x},\bar{y}).$$

This type of assertion has become known as **metric regularity**, and it has a number of important applications in analysis and optimization.

Of course, some hypothesis will be required in order to arrive at such a conclusion. It turns out to be very useful to allow nondifferentiable functions φ in the theory, and the hypothesis is framed with that in mind. Below, the notation $\varphi'_x(x,y;v)$ denotes the directional derivative with respect to the x variable; thus, we have by definition

$$\varphi'_x(x,y;v) = \lim_{t \downarrow 0} \frac{\varphi(x+tv,y) - \varphi(x,y)}{t}.$$

Our missing hypothesis is that φ admits (uniform) decrease directions (in x) at nearby points where it is positive; explicitly:

5.30 Hypothesis. *There is an open neighborhood V of $(\bar{x}, \bar{y})$ and $\delta > 0$ with the following property: at any point $(x, y) \in V$ for which $\varphi(x, y) > 0$, the directional derivatives $\varphi'_x(x, y; v)$ exist for every v and satisfy*

$$\inf_{\|v\| \leqslant 1} \varphi'_x(x, y; v) \leqslant -\delta.$$

5.31 Theorem. *In the presence of* Hypothesis 5.30, *setting $K = 1/\delta$, there exists a neighborhood U of $(\bar{x}, \bar{y})$ such that*

$$d(x, S(y)) \leqslant K\varphi(x, y) \ \forall (x, y) \in U.$$

Proof. It is a consequence of Hypothesis 5.30 that there exist $R > 0$ and a neighborhood W of $\bar{y}$ such that

$$y \in W, \ \|x - \bar{x}\| < R, \ \varphi(x, y) > 0 \implies \inf_{\|v\| \leqslant 1} \varphi'_x(x, y; v) \leqslant -\delta. \qquad (*)$$

We claim that the following implication holds:

$$y \in W, \ \|x - \bar{x}\| + \varphi(x, y)/\delta < R \implies d(x, S(y)) \leqslant \varphi(x, y)/\delta.$$

This evidently yields the theorem, since the conditions on the left are satisfied when (x, y) is sufficiently close to $(\bar{x}, \bar{y})$. We establish the claim by the absurd. If it is false, there exist (x, y) and $\varepsilon > 0$ such that all the following hold:

$$y \in W, \ \|x - \bar{x}\| + \varphi(x, y)/\delta < R, \ d\big(x, S(y)\big) > \varphi(x, y)/\delta + \varepsilon =: r.$$

We may reduce ε as required to further satisfy

$$\|x - \bar{x}\| + \varphi(x, y)/\delta + \varepsilon < R.$$

It follows that $B(x, r) \subset B^{\circ}(\bar{x}, R)$ and that $\varphi(u, y) > 0$ for all $u \in B(x, r)$. We now invoke the decrease principle (Theorem 5.22), which applies to $f := \varphi(\cdot, y)$ on $B(x, r)$ in view of the condition $(*)$ above (as noted in Exer. 5.23). We obtain:

$$\inf\big\{\varphi(u, y) : u \in B^{\circ}(x, r)\big\} \leqslant \varphi(x, y) - \delta r = -\delta\varepsilon < 0,$$

which is the required contradiction (since φ is nonnegative). □

Stability of equations. The remainder of this section obtains various consequences of Theorem 5.31. The first of these was an early precursor of the more general metric regularity presented above; it is useful in dealing with an explicit equation $F(x) = y$. (Below, the notation $F^{-1}(y)$ refers to the set of points x satisfying this equation.) In the theory developed above, this corresponds to taking $\varphi(x, y) = \|F(x) - y\|$.

5.32 Theorem. (Graves-Lyusternik) *Let $F : X \to Y$ be a mapping which is continuously differentiable in a neighborhood of a point $\bar{x}$, and such that $F'(\bar{x})$ is surjective: $F'(\bar{x})X = Y$. Set $\bar{y} = F(\bar{x})$. Then there is a neighborhood U of $(\bar{x}, \bar{y})$ and $K > 0$ such that*

$$d\big(x, F^{-1}(y)\big) \leqslant K \| F(x) - y \|_Y \quad \forall (x,y) \in U.$$

Proof. By the open mapping theorem 5.26, there exists $\delta > 0$ such that

$$F'(\bar{x}) B_X \supset 3\delta B_Y .$$

Furthermore, the continuity of the map $x \mapsto F'(x)$ in a neighborhood of $\bar{x}$ implies the existence of $R > 0$ such that

$$x \in B_X(\bar{x}, R) \implies \big[F'(\bar{x}) - F'(x)\big] B_X \subset \delta B_Y . \tag{1}$$

For such x, we have

$$\begin{aligned} \delta B_Y + 2\delta B_Y = 3\delta B_Y \subset F'(\bar{x}) B_X &= \big\{ [F'(\bar{x}) - F'(x)] + F'(x) \big\} B_X \\ &\subset [F'(\bar{x}) - F'(x)] B_X + F'(x) B_X \subset \delta B_Y + F'(x) B_X \text{ (by (1))}, \end{aligned}$$

which implies $2\delta B_Y \subset \text{cl}\,(F'(x) B_X)$ (see Exer. 2.45). By Theorem 5.24, we obtain

$$x \in B_X(\bar{x}, R) \implies F'(x) B_X^{\circ} \supset \delta B_Y . \tag{2}$$

Now let us define $\varphi(x,y) = \| F(x) - y \|_Y$ (there is no harm in supposing that F is globally defined and continuous, so that φ is as well). Note that $\varphi(\bar{x}, \bar{y}) = 0$. We proceed to prepare an appeal to Theorem 5.31, by verifying Hypothesis 5.30.

Let $x \in B_X(\bar{x}, R)$ and $y \in Y$ be such that $\varphi(x,y) > 0$. By (2), there exists $v \in B_X^{\circ}$ such that

$$F'(x) v = -\delta \, \frac{F(x) - y}{\| F(x) - y \|_Y} .$$

Then $\varphi_x'(x, y; v)$ is given by

$$\begin{aligned} &\lim_{t \downarrow 0} \big\{ \| F(x + tv) - y \| - \| F(x) - y \| \big\} / t \\ &= \lim_{t \downarrow 0} \big\{ \| F(x) + t F'(x) v - y \| - \| F(x) - y \| \big\} / t \text{ (we may neglect } o(t)) \\ &= \lim_{t \downarrow 0} \left\{ \left\| F(x) - y - \delta t \, \frac{F(x) - y}{\| F(x) - y \|} \right\| - \| F(x) - y \| \right\} / t = -\delta. \end{aligned}$$

This verifies Hypothesis 5.30, and allows us to invoke Theorem 5.31, which immediately gives the result (for $K = 1/\delta$). □

5.33 Corollary. *If the mapping $F : X \to Y$ is continuously differentiable in a neighborhood of a point $\bar{x}$, and if $F'(\bar{x})$ is surjective, then, for every $\varepsilon > 0$, the set $F(B(\bar{x}, \varepsilon))$ contains a neighborhood of $F(\bar{x})$.*

5.34 Exercise. Prove the corollary. □

Remark. When $F : \mathbb{R}^n \to \mathbb{R}^k$ is differentiable at $\bar{x}$, we may identify $F'(\bar{x})$ with the linear mapping induced by the $k \times n$ Jacobian matrix $DF(x)$. The surjectivity of the mapping, as we know from linear algebra, is equivalent to this matrix having rank k (which can only happen if $k \leqslant n$).

The tangent space of a Banach manifold. The next use of metric regularity is to give conditions under which the tangent cone to a level set is determined by the derivative of the underlying function. The discussion centered around Exer. 1.40 will now be completed.

5.35 Theorem. *Let X and Y be Banach spaces, and let S be given by*

$$S = \{ u \in X : F(u) = 0 \},$$

where the map $F : X \to Y$ is continuously differentiable near $x \in S$. If $F'(x)$ is surjective, then $T_S(x)$ and $N_S(x)$ are the linear subspaces described as follows:

$$T_S(x) = \{ v \in X : \langle F'(x), v \rangle = 0 \}, \quad N_S(x) = F'(x)^* Y^*$$

and we have

$$T_S(x) = N_S(x)^{\triangle}, \quad N_S(x) = T_S(x)^{\triangle}.$$

The equality for the normal cone in this final assertion always holds, of course, by definition; we are stressing the fact that in this setting, the tangent and normal cones are mutually polar (actually, orthogonal) linear subspaces. In regard to the normal cone, the theorem asserts that an element ζ of $N_S(x)$ is precisely one that can be expressed in the form $F'(x)^* \Lambda$ for some $\Lambda \in Y^*$, where $F'(x)^*$ denotes the adjoint of $F'(x)$; then we have

$$\langle \zeta, u \rangle = \langle F'(x)^* \Lambda, u \rangle = \langle \Lambda, F'(x) u \rangle \;\; \forall u \in X.$$

Proof. First, let $v \in T_S(x)$; we show that $\langle F'(x), v \rangle = 0$. Now v, as a tangent vector, can be expressed in the form $\lim_{i \to \infty} (x_i - x)/t_i$ for sequences $x_i \to x$ (in S) and $t_i \downarrow 0$. Let us set $v_i = (x_i - x)/t_i$. Then

$$\langle F'(x), v \rangle = \lim_{i \to \infty} \frac{F(x + t_i v) - F(x)}{t_i} = \lim_{i \to \infty} \frac{F(x + t_i v_i)}{t_i} = \lim_{i \to \infty} \frac{F(x_i)}{t_i} = 0,$$

where the second equality holds because F is Lipschitz near x, and $F(x) = 0$.

Conversely, let v satisfy $\langle F'(x), v\rangle = 0$; we show that $v \in T_S(x)$. Take any sequence $t_i \downarrow 0$. By Theorem 5.32, there exists, for all i sufficiently large, a point x_i such that

$$F(x_i) = 0,\ \|x + t_i v - x_i\|_X \leqslant K\,\|F(x + t_i v)\|_Y + t_i^2$$

(the last term is there to reflect the fact that the distance $d(x+t_i v, F^{-1}(0))$ may not be attained). Dividing by t_i and letting i tend to ∞ reveals

$$\lim_{i\to\infty} (x_i - x)/t_i = v,$$

which confirms that $v \in T_S(x)$.

We turn now to the formula $T_S(x) = N_S(x)^{\triangle}$. Let us define a convex cone (in fact, subspace) in X^* as follows:

$$\Sigma = F'(x)^* Y^* = \big\{\Lambda \circ F'(x) : \Lambda \in Y^*\big\}.$$

Then

$$v \in \Sigma^{\triangle} \iff \langle \Lambda \circ F'(x), v\rangle \leqslant 0 \ \forall \Lambda \in Y^* \iff \langle F'(x), v\rangle = 0,$$

which shows that the polar of Σ is $T_S(x)$. If Σ is weak* closed, then we have

$$N_S(x) = T_S(x)^{\triangle} = \Sigma^{\triangle\triangle} = \Sigma$$

(see Prop. 4.34), which implies $T_S(x) = N_S(x)^{\triangle}$, the required conclusion.

There remains to verify, however, that Σ is weak* closed. This is simple to do when Y is finite dimensional, but for the general case treated here we call upon Exer. 8.48, according to which it suffices to prove:

Lemma. *The set $\Sigma_B = B_* \cap \Sigma$ is weak* closed.*

By the open mapping theorem, there exists $\delta > 0$ such that $F'(x)B_X \supset \delta B_Y$. It follows that

$$\|\Lambda \circ F'(x)\|_{X^*} \geqslant \delta\,\|\Lambda\|_{Y^*} \ \ \forall \Lambda \in Y^*.$$

This implies that the set

$$\Gamma = \big\{\Lambda \in Y^* : \|\Lambda \circ F'(x)\|_{X^*} \leqslant 1\big\} = \bigcap\nolimits_{v \in B} \big\{\Lambda \in Y^* : \langle \Lambda, F'(x)v\rangle \leqslant 1\big\}$$

is bounded; it follows directly from the way it is defined that it is weak* closed. Then it is weak* compact, by Cor. 3.15.

We claim that the map $T : Y^* \to X^*$ defined by $T\Lambda = \Lambda \circ F'(x)$ is continuous for the weak* topologies of Y^* and X^*. To prove this, it suffices, by Theorem 3.1(e), to verify that, for any $v \in X$, the map $\Lambda \mapsto \langle \Lambda \circ F'(x), v\rangle$ is continuous for $\sigma(Y^*, Y)$. But this is evidently the case, since the map amounts to evaluation of Λ at the point $F'(x)v$. Since Σ_B is the image of the weak* compact set Γ under this continuous map, we deduce that Σ_B is weak* compact, and hence weak* closed. $\square$

Classical manifolds. Let $F:\mathbb{R}^n\to\mathbb{R}^k$ be continuously differentiable, where the integer k satisfies $1\leqslant k<n$. The set $S=F^{-1}(0)$ then corresponds to a classical manifold, when the following **rank condition** is posited:

$$F(x)=0 \implies \operatorname{rank} DF(x)=k,$$

where $DF(x)$ is the $k\times n$ Jacobian matrix.

5.36 Exercise. Prove that the rank condition is equivalent to:

$$F(x)=0,\ \lambda\in\mathbb{R}^k,\ 0=D\langle\lambda,F\rangle(x) \implies \lambda=0. \qquad \square$$

As remarked upon earlier, it is natural to view normal vectors to sets in $\mathbb{R}^n$ as points in $\mathbb{R}^n$ itself, as is done in the next result.

5.37 Corollary. *If the rank condition holds, and if $x\in S$, then $T_S(x)$ and $N_S(x)$ are orthogonal subspaces of $\mathbb{R}^n$ of dimension $n-k$ and k respectively. $N_S(x)$ is the subspace spanned by the k independent vectors $DF^i(x)$ $(i=1,2,\dots,k)$.*

Proof. We know from linear algebra that the surjectivity of $F'(x)$ is equivalent to the rank condition, so Theorem 5.35 applies. We find that $T_S(x)$ is the null space of the matrix $DF(x)$, a vector subspace of $\mathbb{R}^n$ of dimension $n-k$. Then $N_S(x)$ is the orthogonal complement of that subspace, and has the basis described. $\square$

The inverse function theorem. When $X=Y$ and the injectivity of $F'(\bar{x})$ is added to the hypotheses of Theorem 5.32, we obtain a Banach space version of the inverse function theorem. (The proof is outlined in Exer. 8.37.)

5.38 Theorem. *Let $F:X\to X$ be a mapping which is continuously differentiable in a neighborhood of a point $\bar{x}$, and such that $F'(\bar{x})$ is a bijection. Set $\bar{y}=F(\bar{x})$. Then there exist open neighborhoods A of $\bar{x}$ and W of $\bar{y}$ and a continuously differentiable function $\hat{x}:W\to X$ such that*

$$\hat{x}(\bar{y})=\bar{x},\quad F\big(\hat{x}(y)\big)=y\ \ \forall\, y\in W,\quad \hat{x}\big(F(x)\big)=x\ \ \forall\, x\in A.$$

5.39 Example. (Systems of inequalities) Let us consider the stability of the solutions x of a system of *inequalities* $g(x)\leqslant 0$. We remark that classical implicit function methods do not readily apply here.

We are given $g:X\to\mathbb{R}^m$ continuously differentiable. For $y\in\mathbb{R}^m$, we define

$$\Gamma(y)=\big\{x\in X:\, g(x)\leqslant y\big\},$$

where the inequality is understood in the vector sense. In order to use Theorem 5.31, we now require a function φ which equals zero precisely when x belongs to $\Gamma(y)$. Letting $g=(g^1,g^2,\dots,g^m)$, and writing $y=(y^1,y^2,\dots,y^m)$, we choose

$$\varphi(x,y) = \max\{0, g^1(x)-y^1, g^2(x)-y^2, \ldots, g^m(x)-y^m\}.$$

It follows that φ is continuous and nonnegative, and that $\varphi(x,y) = 0$ if and only if $x \in \Gamma(y)$.

Notation: Let the point $\bar{x}$ satisfy $g(\bar{x}) \leqslant 0$. We denote by $I(\bar{x})$ the set of indices i in $\{1,2,\ldots,m\}$ such that $g^i(\bar{x}) = 0$ (if any).

We are ready to state a consequence of Theorem 5.31.

Proposition. *Suppose that the elements* $\{Dg^i(\bar{x}) : i \in I(\bar{x})\}$ *in* X^* *are positively linearly independent. Then there exist* K *and a neighborhood* U *of* $(\bar{x},0)$ *in* $X \times \mathbb{R}^m$ *such that*

$$d\big(x, \Gamma(y)\big) \leqslant K\varphi(x,y), \quad (x,y) \in U.$$

The stated stability property follows directly from Theorem 5.31, provided that Hypothesis 5.30 holds for (x,y) in a neighborhood of $(\bar{x},0)$. This is easily verified with the help of some upcoming results from nonsmooth analysis which characterize the directional derivatives of a nondifferentiable "max function" such as φ. We resume the analysis later, therefore, in Example 10.25. □

5.5 Reflexive spaces and weak compactness

The key to obtaining weak compactness in a normed space X turns out to lie in the fact that X has a clone living in its bidual X^{**}, by which we mean $(X^*)^*$, the dual of the dual. Note that X^{**}, as a dual, is a Banach space by Theorem 5.1.

The **canonical injection** is the linear mapping $J : X \to X^{**}$ defined as follows: for given $x \in X$, we take Jx to be the element of X^{**} determined by the formula

$$\langle Jx, \zeta \rangle = \langle \zeta, x \rangle, \ \ \zeta \in X^*.$$

It is clear that this defines a linear functional on X^*. We calculate

$$\|Jx\|_{X^{**}} = \sup_{\zeta \in B_*} \langle Jx, \zeta \rangle = \sup_{\zeta \in B_*} \langle \zeta, x \rangle = \|x\|_X.$$

Thus, J is norm preserving, and therefore continuous; in fact, J is an isometry from X to the subspace $J(X)$ of X^{**}. Thus the bidual space X^{**} contains a copy $J(X)$ of X (and therefore can only be "bigger" than X).

5.40 Exercise. Prove that the mapping $J : X \to X^{**}$ is continuous when X is equipped with its weak topology and the bidual X^{**} is equipped with the topology $\sigma(X^{**}, X^*)$. □

We say that X is **reflexive** when $JX = X^{**}$. Note that this property is invariant with respect to the choice of equivalent norm on X (since the dual and the bidual remain unchanged).

Since a dual space is always a Banach space (Theorem 5.1), it is immediate from Prop. 5.3 that a reflexive normed space is a Banach space. It also follows that the Cartesian product of two reflexive spaces is reflexive. We shall see later that a closed subspace of a reflexive space is reflexive.

5.41 Exercise. Prove that a normed space X is reflexive if and only if $JB_X = B_{**}$, where B_{**} denotes the closed unit ball in X^{**}. □

5.42 Proposition. *Let X, Y be isometric normed spaces. If Y is reflexive, then X is reflexive.*

Proof. Let $T : X \to Y$ be an isometry. Note that ζ belongs to Y^* if and only if $\zeta \circ T$ belongs to X^*; that is, we have $X^* = Y^* \circ T$ (or $X^* = T^* Y^*$, if we use the adjoint of T to express this fact). Now let $z \in X^{**}$. The goal is to exhibit an element $\bar{x} \in X$ such that $z = J_X(\bar{x})$.

We proceed to define a linear functional f on Y^* by

$$f(\zeta) = \langle z, \zeta \circ T \rangle \;\; \forall\, \zeta \in Y^*.$$

Then we have

$$|f(\zeta)| \leqslant \|z\|_{X^{**}} \|\zeta \circ T\|_{X^*} \leqslant \|z\|_{X^{**}} \|\zeta\|_{Y^*}.$$

It follows that $f \in Y^{**}$, so there exists $\bar{y} \in Y$ such that $J_Y(\bar{y}) = f$. Let $\bar{x} = T^{-1}\bar{y}$. Then, for any $\zeta \in Y^*$, we calculate

$$f(\zeta) = \langle z, \zeta \circ T \rangle = \langle J_Y(\bar{y}), \zeta \rangle = \langle \zeta, \bar{y} \rangle = \langle \zeta \circ T, \bar{x} \rangle.$$

Since $Y^* \circ T = X^*$, this implies $\langle z, \xi \rangle = \langle \xi, \bar{x} \rangle \;\; \forall\, \xi \in X^*$. Thus $z = J_X(\bar{x})$. □

5.43 Proposition. *Let X be a Banach space. Then*

(a) *X is reflexive $\iff$ X^* is reflexive.*

(b) *X is reflexive and separable $\iff$ X^* is reflexive and separable.*

Proof. Let X be reflexive, and fix any $f \in X^{***}$. We define an element $\zeta \in X^*$ by the formula

$$\langle \zeta, x \rangle = \langle f, Jx \rangle, \quad x \in X.$$

We proceed to show that $f = J_* \zeta$, where J_* is the canonical injection of X^* into X^{***}. For this purpose, let θ be any point in X^{**}. Since X is reflexive, there exists $x \in X$ such that $Jx = \theta$. Then we have

$$\langle J_* \zeta, \theta \rangle = \langle \theta, \zeta \rangle = \langle Jx, \zeta \rangle = \langle \zeta, x \rangle = \langle f, Jx \rangle = \langle f, \theta \rangle.$$

Thus $J_* \zeta = f$, and X^* is reflexive.

Now suppose that X^* is reflexive. If $JX \neq X^{**}$, we may separate (see Theorem 2.39) to find $f \in X^{***}$ such that $f(JX) = 0$, $f \neq 0$. (We have used the fact that $J(X)$ is closed by Prop. 5.3.) By hypothesis, we have $f = J_* \zeta$ for some $\zeta \in X^*$. Then, for any $x \in X$,

$$0 = \langle f, Jx \rangle = \langle J_* \zeta, Jx \rangle = \langle Jx, \zeta \rangle = \langle \zeta, x \rangle.$$

There results $\zeta = 0$, whence $f = J_* \zeta = 0$, a contradiction which proves that X is reflexive.

If X is reflexive and separable, then $X^{**} = JX$ is separable, as an isometric image, and consequently X^* is separable (Theorem 3.19) as well as reflexive (by the above). The converse is evident. □

5.44 Exercise. Let X be a Banach space. We set $X^{\{0\}} = X$, and then

$$X^{\{n+1\}} = \left(X^{\{n\}}\right)^*, \ n = 0, 1, 2, \ldots$$

By means of the canonical injections, we may consider that two nondecreasing sequences of successive biduals are obtained: $X^{\{2n\}}$ $(n \geqslant 0)$, and $X^{\{2n+1\}}$ $(n \geqslant 0)$. Prove that only the two following cases arise: either both sequences are strictly increasing, or else both sequences are constant. □

5.45 Proposition. *Let X and Y be normed spaces that admit isometries $\Lambda : X \to Y^*$ and $T : Y \to X^*$ such that*

$$\langle \Lambda x, y \rangle = \langle Ty, x \rangle \ \forall x \in X, \ y \in Y.$$

Then X and Y are reflexive Banach spaces.

Proof. We show that X is reflexive. To this end, let $\theta \in X^{**}$. Then the formula $\psi(y) = \langle \theta, Ty \rangle$ $(y \in Y)$ defines an element ψ of Y^*. Thus, there exists $\bar{x} \in X$ such that $\psi = \Lambda \bar{x}$.

Now let $\zeta \in X^*$. There exists $y \in Y$ such that $Ty = \zeta$. We calculate

$$\langle \theta, \zeta \rangle = \langle \theta, Ty \rangle = \langle \psi, y \rangle = \langle \Lambda \bar{x}, y \rangle = \langle Ty, \bar{x} \rangle \text{ (by hypothesis)} = \langle \zeta, \bar{x} \rangle.$$

Thus, $\langle \theta, \zeta \rangle = \langle \zeta, \bar{x} \rangle \ \forall \zeta \in X^*$; that is, $\theta = J\bar{x}$. □

We shall apply the criterion above to study the reflexivity of the sequence spaces ℓ^p defined in Example 1.6.

5.46 Proposition. *The Banach space ℓ^p $(1 \leqslant p \leqslant \infty)$ is reflexive if and only if $1 < p < \infty$.*

It may be instructive to mention here an *incorrect* way to prove the reflexivity of ℓ^p for $1 < p < \infty$. We know that the dual of ℓ^p is isometric to ℓ^q, where q is the conjugate exponent (see Example 1.27). The dual of ℓ^q, in turn, is isometric to ℓ^p. Thus ℓ^p is isometric to its bidual. It is tempting to assert, simply on the strength of this, that ℓ^p is reflexive. But this reasoning is incorrect, for in the definition of reflexivity, it is specified that the isometry is the canonical injection; it is not enough that X and X^{**} be isometric. (Yes, we assure the reader that there do exist Banach spaces that are isometric to their bidual, but are not reflexive.)

Proof. We know that ℓ^1 is separable (see Example 3.17). If ℓ^1 were reflexive, then $(\ell^1)^*$ would be separable and reflexive, in view of Prop. 5.43. But ℓ^∞ is isometric to $(\ell^1)^*$, so that ℓ^∞ would be separable, which is not the case. We conclude that ℓ^1 is not reflexive. Since ℓ^1 is not reflexive, neither is its dual $(\ell^1)^*$, and therefore, neither is ℓ^∞.

There remains to prove that $X = \ell^p$ is reflexive for $1 < p < \infty$; we do so with the help of Prop. 5.45. Let q be the exponent conjugate to p, and set $Y = \ell^q$. We have an isometry $T : \ell^q \to (\ell^p)^*$ which operates as follows: let $g \in \ell^q$; then Tg is the element of $(\ell^p)^*$ such that

$$\langle Tg, f\rangle = \sum_{i\geqslant 1} f_i\, g_i \;\; \forall f \in \ell^p.$$

Similarly, there exists $\Lambda : \ell^p \to (\ell^q)^*$ such that

$$\langle \Lambda f, g\rangle = \sum_{i\geqslant 1} f_i\, g_i \;\; \forall g \in \ell^q.$$

Then $\langle \Lambda f, g\rangle = \langle Tg, f\rangle \;\; \forall f \in \ell^p,\ g \in \ell^q$, and it follows from Prop. 5.45 that ℓ^p is reflexive. □

Weak compactness. In $\mathbb{R}^n$, the fact that closed bounded sets are compact is a highly useful one, as the reader knows. We prove below that reflexive spaces enjoy a somewhat similar property: a closed, bounded, *convex* subset of a reflexive Banach space is weakly compact. The key is the following.

5.47 Theorem. *A Banach space X is reflexive if and only if its closed unit ball is weakly compact.*

Proof. Suppose first that X is reflexive; thus $JB = B_{**}$ (see Exer. 5.41). Let $\{V_\alpha\}$ be a covering of B by weakly open sets. We prove the existence of a finite subcover. Without loss of generality, we may suppose that each V_α is the trace on the ball of a canonical base element for the weak topology:

$$V_\alpha = \bigcap_{i\in F_\alpha} \big\{ x \in X,\ \|x\| \leqslant 1 : \big|\langle \zeta_{\alpha,i}, x - x_\alpha\rangle\big| < r_\alpha \big\},$$

where $\{\, \zeta_{\alpha,i} : i \in F_\alpha \,\}$ is a finite collection in X^* for each α, and where $x_\alpha \in X$ and $r_\alpha > 0$. We observe that the sets

$$\begin{aligned} JV_\alpha &= \bigcap_{i\in F_\alpha} \left\{ Jx \in X^{**},\ \|Jx\| \leqslant 1 : \left|\langle Jx - Jx_\alpha\,, \zeta_{\alpha,i}\rangle\right| < r_\alpha \right\} \\ &= \bigcap_{i\in F_\alpha} \left\{ \theta \in X^{**},\ \|\theta\| \leqslant 1 : \left|\langle \theta - \theta_\alpha\,, \zeta_{\alpha,i}\rangle\right| < r_\alpha \right\} \end{aligned}$$

constitute a covering of $JB = B_{**}$, where $\theta_\alpha := Jx_\alpha$. (The reflexivity of X is essential here.) Further, these sets are the trace on the ball B_{**} of sets which are open for the topology $\sigma(X^{**}, X^*)$ (by the very definition of this topology). But in the space X^{**} equipped with this topology, B_{**} is compact (Cor. 3.15). We may therefore extract a finite subcover, which leads to the required subcover of $\{V_\alpha\}$ in an evident way.

The converse uses the following result known as Goldstine's lemma:

Lemma. *For the topology* $\sigma(X^{**}, X^*)$, $J(B)$ *is dense in* B_{**}.

To see this, fix $\theta \in B_{**}$ and let V be any neighborhood of θ in the topology $\sigma(X^{**}, X^*)$. We desire to prove that $V \cap J(B) \neq \emptyset$. Without loss of generality, we may suppose that V has the canonical form

$$V = \bigcap_{i=1}^{n} \left\{ \theta' \in X^{**} : \left|\langle \theta' - \theta, \zeta_i\rangle\right| < r \right\},$$

where $\zeta_i \in X^*$ and $r > 0$. Thus we seek to prove the existence of $x \in B$ such that

$$\left|\langle Jx, \zeta_i\rangle - \langle \theta, \zeta_i\rangle\right| = \left|\langle \zeta_i, x\rangle - \langle \theta, \zeta_i\rangle\right| < r,\ i = 1, 2, \dots, n.$$

Let us define $\varphi : X \to \mathbb{R}^n$ by $\varphi(x) = (\langle \zeta_1, x\rangle, \langle \zeta_2, x\rangle, \dots, \langle \zeta_n, x\rangle)$, and set

$$v = \big(\langle \theta, \zeta_1\rangle, \langle \theta, \zeta_2\rangle, \dots, \langle \theta, \zeta_n\rangle\big) \in \mathbb{R}^n.$$

It suffices to show that $v \in \operatorname{cl}\varphi(B)$. If such is not the case, we may invoke the strict case of the separation theorem 2.37 in order to deduce the existence of $\beta \in \mathbb{R}^n$ and $\gamma \in \mathbb{R}$ such that

$$\varphi(x) \bullet \beta < \gamma < v \bullet \beta \quad \forall x \in B.$$

Thus, for all $x \in B$, we have

$$\begin{aligned} \Big\langle \sum_1^n \beta_i\,\zeta_i, x\Big\rangle &= \sum_1^n \beta_i \langle \zeta_i, x\rangle = \varphi(x) \bullet \beta < \gamma < v \bullet \beta \\ &= \sum_1^n \beta_i \langle \theta, \zeta_i\rangle = \Big\langle \theta, \sum_1^n \beta_i\,\zeta_i\Big\rangle \leqslant \Big\| \sum_1^n \beta_i\,\zeta_i \Big\|_*, \end{aligned}$$

since $\|\theta\|_{**} \leqslant 1$. Taking the supremum over $x \in B$ on the left leads to a contradiction, which completes the proof of the lemma.

Suppose now that the ball B is weakly compact. Then JB is compact in X^{**} equipped with the topology $\sigma(X^{**}, X^*)$ (by Exer. 5.40). In a Hausdorff topology such as $\sigma(X^{**}, X^*)$, a compact set is also closed. Then, by the lemma, we deduce $JB = B_{**}$, which is equivalent to the reflexivity of X. □

5.48 Corollary. *Any closed, bounded, convex subset A of a reflexive Banach space is weakly compact.*

5.49 Exercise. Prove that a closed subspace L of a reflexive Banach space X is reflexive. □

The weak topology is never metrizable when the underlying normed space is infinite-dimensional (Exer. 8.43). In certain non metrizable spaces, there is a distinction between sets which are compact and those which are *sequentially* compact. It turns out that this potential distinction does not arise in the current context:

5.50 Theorem. *Any closed, bounded, convex subset A of a reflexive Banach space is weakly sequentially compact.*

Proof. The proof is short, but it manages to provide a review of many prior results. Let x_i be a sequence in A. We prove the existence of $x \in A$ and a subsequence x_{n_i} converging to x.

For some $r > 0$, all the terms x_i of the sequence are contained in the ball rB. Let L be the closed subspace of X generated by the sequence x_i. Then L is separable (see the argument given in the proof of Theorem 3.19). Furthermore, L is reflexive by Exer. 5.49. It follows that the closed ball B_L is weakly compact in L, by Theorem 5.47, as is rB_L.

We know that L^* is separable by Prop. 5.43. Then the weak topology of L, when restricted to rB_L, is metrizable, by Theorem 3.20. This implies weak sequential compactness of rB_L, whence the existence of a subsequence x_{n_i} converging weakly in L to a limit x. Thus, we have

$$\langle \zeta, x_{n_i} \rangle \to \langle \zeta, x \rangle \;\; \forall \zeta \in L^*.$$

But (by Cor. 1.34) this is equivalent to the weak convergence in X of x_{n_i} to x. Finally, the point x belongs to A, since A is weakly closed, being closed and convex (Theorem 3.6). □

The Direct Method. If this were a multimedia presentation, the reader might hear a drumroll at this point, and possibly trumpets, for we are on the brink of attaining a long-sought goal: to formulate a general existence theorem for minimization. Several of the tools we have studied (lower semicontinuity, convexity, weak compactness) were invented for precisely such a purpose.

The problem we consider is the familiar one:

$$\text{Minimize } f(x) \text{ subject to } x \in A. \tag{P}$$

A solution of (P) refers, of course, to a point $\bar{x} \in A$ such that $f(\bar{x}) = \inf_A f$.

5.51 Theorem. *Under the following hypotheses,* $\inf_A f$ *is finite, and the problem* (P) *admits a solution*:

(a) *A is a closed convex subset of a reflexive Banach space X*;

(b) *$f : X \to \mathbb{R}_\infty$ is convex and lower semicontinuous, and* $\operatorname{dom} f \cap A \neq \emptyset$;

(c) *For every $M \in \mathbb{R}$, the set $\{x \in A : f(x) \leqslant M\}$ is bounded.*

In fact, any minimizing sequence x_n for the problem (P) *admits a subsequence converging weakly to a solution $\bar{x}$ of the problem.*

The proof is carried out by invoking weak compactness and lower semicontinuity in connection with a minimizing sequence, a technique that has become known as the **direct method**, following its introduction under this name by Tonelli, in the context of the calculus of variations.

Proof. Since $A \cap \operatorname{dom} f \neq \emptyset$, there exists a point $x_0 \in A$ such that $f(x_0) < \infty$. Thus $\inf_A f$ is either finite or $-\infty$. Let x_n be any minimizing sequence; that is, a sequence of points in A such that

$$f(x_n) \to \inf_A f < f(x_0) + 1 .$$

For n sufficiently large, we have $f(x_n) \leqslant f(x_0) + 1$, so that, by the growth condition (c), the sequence x_n is bounded. Since X is reflexive, there exists by Theorem 5.50 a subsequence x_{n_i} of x_n converging weakly to a limit $\bar{x} \in X$.

Because A is convex and closed, A is weakly closed (Theorem 3.6), whence $\bar{x} \in A$. Because f is lsc and convex, f is weakly lsc (Cor. 3.7), whence

$$f(\bar{x}) \leqslant \lim_{i \to \infty} f\big(x_{n_i}\big) = \inf_A f .$$

It follows that $\inf_A f$ is finite and attained at $\bar{x}$, which solves the problem (P). □

The growth condition. We remark that the growth condition (c) used in the theorem, a joint one involving both f and A, can result in certain cases from a property of either the set A or the function f alone. The following cases are notable ones implying (c):

- The set A is bounded, or
- The function f is *coercive*: $\lim_{\|x\| \to \infty} f(x) = \infty$.

5.52 Exercise. Let S be a closed, convex, nonempty subset of a Banach space X.

(a) If X is reflexive, show that for each $x \in X$ there exists at least one closest point in S to x; that is, a point $p \in S$ such that $d_S(x) = \|x - p\|$. Deduce, in particular, that S contains an element of least norm.

(b) Let X be the Banach space $C[0,1]$, and let S be the set of points $f \in X$ such that

$$\int_0^{1/2} f(t)\,dt - \int_{1/2}^1 f(t)\,dt = 1.$$

Prove that S is nonempty, closed, and convex, but that S does not possess an element of least norm. Deduce from this that $C[0,1]$ is not reflexive. □

The most interesting applications of the direct method lie ahead, but the reader will find that the salient points nonetheless emerge in the following illustration.

5.53 Exercise. Let $f_i : \mathbb{R} \to \mathbb{R}_+$ be a sequence of convex functions such that

$$\textstyle\sum_{i=1}^{\infty} f_i(0) < +\infty.$$

For given $r \in (1,\infty)$, define $f : \ell^r \to \mathbb{R}_\infty$ by

$$f(x) = \textstyle\sum_{i=1}^{\infty} f_i(x_i),$$

and set

$$A = \left\{ x \in \ell^r : 0 \leqslant x_i \ \forall i, \ \textstyle\sum_{i=1}^{\infty} x_i \leqslant 1 \right\}.$$

Prove that $\inf_A f$ is finite and attained. Show that this assertion *fails* for $r = 1$. □

Chapter 6
Lebesgue spaces

The Lebesgue spaces $L^p(\Omega)$ play a central role in many applications of functional analysis. This chapter focuses upon their basic properties, as well as certain attendant issues that will be important later. Notable among these are the semicontinuity of integral functionals, and the existence of measurable selections.

6.1 Uniform convexity and duality

We begin by identifying a geometric property of the norm which, when present, turns out to have a surprising consequence. Let X be a normed space.

6.1 Definition. *X is* **uniformly convex** *if it satisfies the following property*:

$$\forall\, \varepsilon > 0,\ \exists\, \delta > 0 \text{ such that } x \in B,\, y \in B,\ \|x-y\| > \varepsilon \implies \left\| \frac{x+y}{2} \right\| < 1-\delta\,.$$

In geometric terms, this is a way of saying that the unit ball is curved.[1] The property depends upon the choice of the norm on X, even among equivalent norms, as one can see even in $\mathbb{R}^2$.

6.2 Exercise. The following three norms on $\mathbb{R}^2$ are equivalent:

$$\|(x,y)\|_1 = |x|+|y|, \quad \|(x,y)\|_2 = |(x,y)| = \big\{|x|^2+|y|^2\big\}^{1/2},$$
$$\|(x,y)\|_\infty = \max\big(|x|,|y|\big)\,.$$

Which ones make $\mathbb{R}^2$ a uniformly convex normed space? □

[1] It turns out that the ball in $\mathbb{R}$ is curved in this sense, although it may seem rather straight to the reader.

F. Clarke, *Functional Analysis, Calculus of Variations and Optimal Control*,
Graduate Texts in Mathematics 264, DOI 10.1007/978-1-4471-4820-3_6,

Despite the fact that uniform convexity is a norm-dependent property, the very existence of such a norm yields an intrinsic property of the underlying space, one that does not depend on the choice of equivalent norm.

6.3 Theorem. (Milman) *Any uniformly convex Banach space is reflexive.*

Proof. Let $\theta \in X^{**}$ satisfy $\|\theta\|_{**} = 1$, and fix any $\varepsilon > 0$. We shall prove the existence of $x \in B$ such that $\|Jx - \theta\|_{**} \leqslant \varepsilon$. Since JB is closed in X^{**} (see Prop. 5.3), this implies $JB = B_{**}$, and consequently that $JX = X^{**}$, so that X is reflexive.

Let δ correspond to ε as in the definition of uniform convexity. We choose $\zeta \in X^*$, $\|\zeta\|_* = 1$, such that $\langle\theta, \zeta\rangle > 1 - \delta/2$, and we set

$$V = \big\{\theta' \in X^{**} : \big|\langle\theta' - \theta, \zeta\rangle\big| < \delta/2\big\},$$

which is an open neighborhood of θ in the topology $\sigma\big(X^{**}, X^*\big)$. By Goldstine's lemma (see the proof of Theorem 5.47), V intersects JB: there exists $x \in B$ such that

$$\big|\langle\zeta, x\rangle - \langle\theta, \zeta\rangle\big| = \big|\langle Jx - \theta, \zeta\rangle\big| < \delta/2.$$

We claim that $\|Jx - \theta\|_{**} \leqslant \varepsilon$. We reason from the absurd, by supposing that θ lies in W, where W is the complement in X^{**} of the set $Jx + \varepsilon B_{**}$.

Since $Jx + \varepsilon B_{**}$ is closed in $\sigma\big(X^{**}, X^*\big)$, W is open in $\sigma\big(X^{**}, X^*\big)$. Thus $V \cap W$ is an open neighborhood of θ in this topology. By Goldstine's lemma, there exists $y \in B$ such that $Jy \in V \cap W$. Thus we have $|\langle\zeta, y\rangle - \langle\theta, \zeta\rangle| < \delta/2$ by definition of V. We calculate

$$\begin{aligned} 1 - \delta/2 < \langle\theta, \zeta\rangle &= \tfrac{1}{2}\big\{\langle\theta, \zeta\rangle - \langle\zeta, y\rangle\big\} + \tfrac{1}{2}\big\{\langle\theta, \zeta\rangle - \langle\zeta, x\rangle\big\} + \tfrac{1}{2}\langle\zeta, x + y\rangle \\ &< \delta/4 + \delta/4 + \|x + y\|/2. \end{aligned}$$

It follows that $\|x + y\|/2 > 1 - \delta$, whence $\|x - y\| \leqslant \varepsilon$ (from uniform convexity). However, $Jy \in W$ yields $\varepsilon < \|Jy - Jx\| = \|y - x\|$, a contradiction which completes the proof. □

There exist reflexive spaces which fail to admit an equivalent norm that is uniformly convex; thus, the existence of such a norm is not a necessary condition for reflexivity. But it is a useful sufficient condition, notably in the study of the Lebesgue spaces introduced in Example 1.9.

6.4 Theorem. *If $1 < p < \infty$, the Banach space $L^p(\Omega)$ is reflexive.*

Proof. We treat first[2] the case $2 \leqslant p < \infty$. Then, we claim, $L^p(\Omega)$ is uniformly convex, and therefore reflexive by Theorem 6.3.

[2] We follow Brézis [8, théorème IV.10].

When $p \geqslant 2$, it is easy to show (by examining its derivative) that the function

$$\theta(t) = \left(t^2+1\right)^{p/2} - t^p - 1$$

is increasing on $[0,\infty)$, which implies, by writing $\theta(0) \leqslant \theta(\alpha/\beta)$, the inequality

$$\alpha^p + \beta^p \leqslant \left(\alpha^2+\beta^2\right)^{p/2} \quad \forall\, \alpha, \beta \geqslant 0.$$

Now let $a, b \in \mathbb{R}$ and take $\alpha = |a+b|/2,\ \beta = |a-b|/2$; we find

$$\left|\frac{a+b}{2}\right|^p + \left|\frac{a-b}{2}\right|^p \leqslant \left(\left|\frac{a+b}{2}\right|^2 + \left|\frac{a-b}{2}\right|^2\right)^{p/2} = \left(\frac{a^2}{2}+\frac{b^2}{2}\right)^{p/2} \leqslant \frac{a^p}{2}+\frac{b^p}{2}$$

(the last estimate uses the convexity of the function $t \mapsto |t|^{p/2}$, which holds because $p \geqslant 2$). This yields *Clarkson's inequality*:

$$\left\|\frac{f+g}{2}\right\|_{L^p}^p + \left\|\frac{f-g}{2}\right\|_{L^p}^p \leqslant \frac{1}{2}\left(\|f\|_{L^p}^p + \|g\|_{L^p}^p\right) \quad \forall\, f, g \in L^p(\Omega).$$

Fix $\varepsilon > 0$, and suppose that f, g in the unit ball of $L^p(\Omega)$ satisfy $\|f-g\|_{L^p} > \varepsilon$. From Clarkson's inequality we deduce

$$\left\|\frac{f+g}{2}\right\|_{L^p}^p < 1 - \left(\frac{\varepsilon}{2}\right)^p \implies \left\|\frac{f+g}{2}\right\|_{L^p} < 1-\delta,$$

where $\delta = 1 - [1-(\varepsilon/2)^p]^{1/p}$. This verifies the uniform convexity, and completes the proof of the case $p \geqslant 2$.

We now prove that $L^p(\Omega)$ is reflexive for $1 < p < 2$. Let $q = p_*$, and consider the operator $T : L^p(\Omega) \to L^q(\Omega)^*$ defined as follows: for $u \in L^p(\Omega)$, the effect of Tu on $L^q(\Omega)$ is given by

$$\langle Tu, g\rangle = \int_\Omega u(x)\, g(x)\, dx \quad \forall\, g \in L^q(\Omega).$$

Then we have (see Exer. 1.31)

$$\|Tu\|_{L^q(\Omega)^*} = \|u\|_{L^p(\Omega)}.$$

Thus T is an isometry between $L^p(\Omega)$ and a *closed* subspace of $L^q(\Omega)^*$ (since $L^p(\Omega)$ is complete, see Prop. 5.3). Now $q > 2$, so $L^q(\Omega)$ is reflexive (by the case of the theorem proved above); thus, its dual $L^q(\Omega)^*$ is reflexive (Prop. 5.43). Then $T(L^p(\Omega))$, as a closed subspace, is reflexive (Exer. 5.49), and therefore $L^p(\Omega)$ is reflexive as well (Prop. 5.42). □

6.5 Corollary. *The spaces $AC^p[a,b]$ are reflexive for $1 < p < \infty$.*

6.6 Exercise. Let $\Lambda : [0,1]\times\mathbb{R}\times\mathbb{R} \to \mathbb{R}$ be a continuous function having the property that, for almost every $t \in [0,1]$, the function $(x,v) \mapsto \Lambda(t,x,v)$ is convex (see Example 2.30). We suppose in addition that, for certain numbers $r > 1$, $\alpha > 0$ and β, we have

$$\Lambda(t,x,v) \geqslant \alpha|v|^r + \beta \;\; \forall (t,x,v) \in [0,1]\times\mathbb{R}\times\mathbb{R}.$$

Fix $x_0, x_1 \in \mathbb{R}$, and consider the following minimization problem (P):

$$\min \;\; f(x) = \int_0^1 \Lambda\big(t, x(t), x'(t)\big)\, dt \;\; : x \in \mathrm{AC}[0,1],\, x(0) = x_0,\, x(1) = x_1.$$

Prove that (P) admits a solution. □

6.7 Exercise. Let $1 < r < \infty$, and let x_i be a bounded sequence of functions in $\mathrm{AC}^r[a,b]$. Prove the existence of $x_* \in \mathrm{AC}^r[a,b]$ and a subsequence x_{i_j} such that

$$x_{i_j} \to x_* \text{ uniformly on } [a,b], \quad {x_{i_j}}' \to x_*' \text{ weakly in } L^r(a,b). \qquad \square$$

6.8 Theorem. (Riesz) *For $1 < p < \infty$, the dual space of $L^p(\Omega)$ is isometric to $L^q(\Omega)$, where q is the conjugate exponent of p. More precisely, each ζ of the dual admits a function $g \in L^q(\Omega)$ (necessarily unique) such that*

$$\langle \zeta, f \rangle = \int_\Omega f(x)\, g(x)\, dx \;\; \forall f \in L^p(\Omega).$$

We then have $\|\zeta\|_{L^p(\Omega)^*} = \|g\|_{L^q(\Omega)}$.

Proof. Let the linear mapping $T : L^q(\Omega) \to L^p(\Omega)^*$ be defined by

$$\langle Tg, f \rangle = \int_\Omega f(x)\, g(x)\, dx \;\; \forall f \in L^p(\Omega).$$

Then, as we know, $\|Tg\|_{L^p(\Omega)^*} = \|g\|_{L^q(\Omega)}$ (see Exer. 1.31), so that T is injective. We proceed to prove that T is surjective, which implies the theorem.

Since $T(L^q(\Omega))$ is closed (as the image of a Banach space under an isometry, see Prop. 5.3), it suffices to prove that $T(L^q(\Omega))$ is dense in $L^p(\Omega)^*$. To prove this, it suffices in turn to prove that (see Theorem 2.39)

$$\theta \in L^p(\Omega)^{**},\; \langle \theta, Tg \rangle = 0 \;\; \forall g \in L^q(\Omega) \implies \theta = 0.$$

We proceed to establish this now. Since $L^p(\Omega)$ is reflexive, there exists $f \in L^p(\Omega)$ such that $\theta = Jf$. Then

$$\langle \theta, Tg \rangle = 0 = \langle Jf, Tg \rangle = \langle Tg, f \rangle = \int_\Omega f(x)\, g(x)\, dx \;\; \forall g \in L^q(\Omega).$$

We discover $f = 0$, by taking $g = |f|^{p-2} f$ (which lies in $L^q(\Omega)$) in the preceding relation, whence $\theta = 0$. □

6.9 Exercise. Characterize the dual of $\mathrm{AC}^p[a,b]$ for $1 < p < \infty$. □

We now proceed to characterize the dual of $L^1(\Omega)$; the proof can no longer rely on reflexivity, however.

6.10 Theorem. *The dual of $L^1(\Omega)$ is isometric to $L^\infty(\Omega)$. More precisely, ζ belongs to $L^1(\Omega)^*$ if and only if there exists $z \in L^\infty(\Omega)$ (necessarily unique) such that*

$$\langle \zeta, f \rangle = \int_\Omega z(x) f(x)\, dx \quad \forall f \in L^1(\Omega).$$

When this holds we have $\|\zeta\|_{L^1(\Omega)^*} = \|z\|_{L^\infty(\Omega)}$.

Proof. That any $z \in L^\infty(\Omega)$ can be used as indicated to engender an element ζ in the dual of $L^1(\Omega)$ is clear, since

$$\langle \zeta, f \rangle \leqslant \|z\|_{L^\infty(\Omega)} \|f\|_{L^1(\Omega)}.$$

Thus any ζ defined in this way satisfies $\|\zeta\|_{L^1(\Omega)^*} \leqslant \|z\|_{L^\infty(\Omega)}$. Let us prove the opposite inequality, for which we may limit attention to the case $\|z\|_{L^\infty(\Omega)} > 0$. For any $\varepsilon > 0$, there exists a measurable subset $S \subset \Omega$ of positive finite measure such that

$$|z(x)| \geqslant \|z\|_{L^\infty(\Omega)} - \varepsilon, \ \ x \in S \text{ a.e.}$$

Set $f(x) = z(x)/|z(x)|$ for $x \in S$, and $f = 0$ elsewhere. Then $f \in L^1(\Omega)$, and we find

$$\langle \zeta, f \rangle = \int_\Omega z(x) f(x)\, dx \geqslant \big(\|z\|_{L^\infty(\Omega)} - \varepsilon\big) \operatorname{meas} S = \big(\|z\|_{L^\infty(\Omega)} - \varepsilon\big) \|f\|_{L^1(\Omega)}.$$

It follows that

$$\|\zeta\|_{L^1(\Omega)^*} \geqslant \|z\|_{L^\infty(\Omega)} - \varepsilon.$$

Since $\varepsilon > 0$ is otherwise arbitrary, the assertion concerning $\|\zeta\|_{L^1(\Omega)^*}$ is proved.

There remains to show that every $\zeta \in L^1(\Omega)^*$ is generated by some z as above. We prove this first under the additional hypothesis that Ω is bounded.

Let $\zeta \in L^1(\Omega)^*$. Any $f \in L^2(\Omega)$ belongs to $L^1(\Omega)$, since Ω is bounded; by Hölder's inequality we have:

$$\langle \zeta, f \rangle \leqslant \|\zeta\|_{L^1(\Omega)^*} \|f\|_{L^1(\Omega)} \leqslant \|\zeta\|_{L^1(\Omega)^*} \big[\operatorname{meas}(\Omega)\big]^{1/2} \|f\|_{L^2(\Omega)}.$$

It follows that ζ can be viewed as an element of $L^2(\Omega)^*$. According to Theorem 6.8, there is a unique z in $L^2(\Omega)$ such that (by the preceding inequality)

$$\langle \zeta, f \rangle = \int_\Omega z(x) f(x)\, dx \leqslant \|\zeta\|_{L^1(\Omega)^*} \|f\|_{L^1(\Omega)} \quad \forall f \in L^2(\Omega).$$

Thus, for any $f \in L^\infty(\Omega) \subset L^2(\Omega)$, we have (by rewriting):

$$\int_\Omega \{ \| \zeta \|_{L^1(\Omega)^*} |f(x)| - z(x) f(x) \} \, dx \geqslant 0 .$$

This implies (here we must beg the reader's pardon for a regrettable forward reference: see Theorem 6.32)

$$\| \zeta \|_{L^1(\Omega)^*} |f| - z(x) f \geqslant 0 \;\; \forall f \in \mathbb{R}, \;\; x \in \Omega \text{ a.e.},$$

which yields $|z(x)| \leqslant \| \zeta \|_{L^1(\Omega)^*}$ a.e. Thus z belongs to $L^\infty(\Omega)$, and satisfies

$$\langle \zeta, f \rangle = \int_\Omega z(x) f(x) \, dx \;\; \forall f \in L^\infty(\Omega) .$$

Given any $f \in L^1(\Omega)$, there is a sequence $f_i \in L^\infty(\Omega)$ such that

$$\| f - f_i \|_{L^1(\Omega)} \to 0 .$$

For instance, let $f_i(x) = f(x)$ if $|f(x)| \leqslant i$, and 0 otherwise. We have, by the above

$$\langle \zeta, f_i \rangle = \int_\Omega z(x) f_i(x) \, dx \;\; \forall i \geqslant 1 .$$

Recalling that ζ is continuous, and passing to the limit, we obtain the same conclusion for f; it follows that z represents ζ on $L^1(\Omega)$, as we wished to show. That z is the *unique* function doing this is left as an exercise.

There remains to treat the case in which Ω is unbounded. Let $\zeta \in L^1(\Omega)^*$. For any sufficiently large positive integer k, the set $\Omega_k := \Omega \cap B^\circ(0,k)$ is nonempty. Then ζ induces an element of $L^1(\Omega_k)^*$: we simply extend to Ω any function $f \in L^1(\Omega_k)$ by setting it equal to 0 on $\Omega \setminus \Omega_k$, then apply ζ to the extension. By the above, there is a function $z_k \in L^\infty(\Omega_k)$ such that

$$\langle \zeta, f \rangle = \int_{\Omega_k} z_k(x) f(x) \, dx \;\; \forall f \in L^1(\Omega_k), \;\; \| z_k \|_{L^\infty(\Omega_k)} = \| \zeta \|_{L^1(\Omega_k)^*} \leqslant \| \zeta \|_{L^1(\Omega)^*} .$$

It is clear that each of the functions z_k is necessarily an extension of the preceding ones (by uniqueness), so they define a function $z \in L^\infty(\Omega)$. We claim that this z represents ζ as required. Let f be any element of $L^1(\Omega)$, and let f_k be the function which agrees with f on Ω_k and which is zero on $\Omega \setminus \Omega_k$. Then

$$\langle \zeta, f_k \rangle = \int_{\Omega_k} z(x) f_k(x) \, dx = \int_\Omega z(x) f_k(x) \, dx .$$

But $f_k \to f$ in $L^1(\Omega)$, so in the limit we obtain

$$\langle \zeta, f \rangle = \int_\Omega z(x) f(x) \, dx .$$

□

6.11 Exercise. Let f_i be a sequence in $L^\infty(0,1)$ such that, for each $g \in L^1(0,1)$, we have

$$\inf_{i \geqslant 1} \int_0^1 f_i(t)\, g(t)\, dt > -\infty.$$

Prove the existence of M such that $\| f_i \|_{L^\infty(0,1)} \leqslant M \;\; \forall\, i$. □

6.12 Exercise. Let θ belong to $L^\infty(0,1)$. Prove the existence of a solution to the following minimization problem:

$$\min_{v \in L^1(0,1)} \int_0^1 e^{[(v(t)-1)^2]}\, dt \quad \text{subject to} \quad \int_0^1 \theta(t)\, v(t)\, dt = 0.$$

□

6.13 Proposition. *The spaces $L^1(\Omega)$ and $L^\infty(\Omega)$ are not reflexive.*

Proof. For ease of exposition, as they say, let us suppose that Ω contains a ball $B(0,r)$. We define a function z in $L^\infty(\Omega)$ as follows:

$$z(x) = \begin{cases} 1 - 2^{-n} & \text{if } 2^{-n-1} r \leqslant |x| < 2^{-n} r, \;\; n = 0, 1, 2 \dots \\ 0 & \text{otherwise.} \end{cases}$$

Note that $\| z \|_{L^\infty(\Omega)} = 1$. If $f \neq 0$ is any nonnegative function in $L^1(\Omega)$, then

$$\int_\Omega f(x)\, z(x)\, dx < \int_\Omega f(x)\, dx = \| f \|_{L^1(\Omega)}.$$

If we denote by ζ the element of $L^1(\Omega)^*$ corresponding to z as in Theorem 6.10, it follows that the supremum

$$\sup \big\{ \langle \zeta, f \rangle : \| f \|_{L^1(\Omega)} \leqslant 1 \big\} = \| \zeta \|_{L^1(\Omega)^*}$$

is *not* attained. But if the unit ball in $L^1(\Omega)$ were weakly compact, this supremum would be attained. We deduce from Theorem 5.47 that $L^1(\Omega)$ is not reflexive. It then follows from Theorem 6.10 and Prop. 5.42 that $L^\infty(\Omega)$ is not reflexive. □

We examine next the separability or otherwise of the Lebesgue spaces.

6.14 Proposition. *$L^p(\Omega)$ is separable for $1 \leqslant p < \infty$.*

Proof. We sketch the proof in the case $p = 1$, the other cases being much the same. We also take Ω bounded, a reduction that is easy to justify. Let $f \in L^1(\Omega)$, and let f_i be the function which coincides with f when $|f| \leqslant i$, and which equals 0 otherwise. Then f_i is measurable, and it follows that $f_i \to f$ in $L^1(\Omega)$. By Lusin's

theorem[3] there exists a continuous function g_i on Ω having compact support, which is bounded in absolute value by i, and which agrees with f_i except on a set S_i of measure less than $1/i^2$. Then

$$\int_\Omega |f_i - g_i|\,dx = \int_{\Omega\setminus S_i} |f_i - g_i|\,dx + \int_{S_i} |f_i - g_i|\,dx \leqslant (2i)/i^2 \to 0.$$

Since $f_i \to f$, we deduce that $C(\overline{\Omega})$ is dense in $L^1(\Omega)$. However, the set of polynomials with rational coefficients is dense in $C(\overline{\Omega})$, by the Weierstrass approximation theorem, whence the separability of $L^1(\Omega)$. □

The proof shows that $C_c(\Omega)$, the continuous functions on Ω having compact support in Ω, is dense in $L^1(\Omega)$. It can be shown that $C_c^\infty(\Omega)$ has the same property.

6.15 Exercise. Prove that $L^\infty(\Omega)$ is not separable. □

Weak compactness without reflexivity. Certain useful compactness properties do hold in $L^1(\Omega)$ and $L^\infty(\Omega)$, despite the fact that these spaces fail to be reflexive. We identify two such cases below, in each of which the separability of $L^1(\Omega)$ plays a role.

6.16 Exercise. Let f_i be a bounded sequence in $L^\infty(\Omega)$. Prove the existence of a subsequence f_{i_j} and $f \in L^\infty(\Omega)$ such that

$$g \in L^1(\Omega) \implies \int_\Omega g(x)\, f_{i_j}(x)\,dx \to \int_\Omega g(x)\, f(x)\,dx.$$ □

In applications to come, the reader will find that it is common to deal with a sequence of functions f_i in $L^1(0,T)$ satisfying a uniform bound of the type $|f_i(t)| \leqslant k(t)$ a.e., where k is summable. The following establishes a sequential compactness result that applies to such a situation.

6.17 Proposition. *Let $k(\cdot) \in L^1(\Omega)$, where Ω is an open subset of $\mathbb{R}^n$. Then the set*

$$K = \big\{ f \in L^1(\Omega) : |f(x)| \leqslant k(x),\ x \in \Omega \text{ a.e.} \big\}$$

is weakly compact and sequentially weakly compact in $L^1(\Omega)$.

Proof. Let us set

$$X = L^\infty(\Omega) \text{ equipped with the weak}^* \text{ topology } \sigma\big(L^\infty(\Omega), L^1(\Omega)\big),$$

$$Y = L^1(\Omega) \text{ equipped with the weak topology } \sigma\big(L^1(\Omega), L^\infty(\Omega)\big).$$

---

[3] Let Ω be a bounded open subset of $\mathbb{R}^n$, and let $\varphi : \Omega \to \mathbb{R}$ be measurable, $|\varphi(x)| \leqslant M$ a.e. For every $\varepsilon > 0$ there exists $g : \Omega \to \mathbb{R}$, continuous with compact support, having $\sup_\Omega |g| \leqslant M$, such that meas$\{x \in \Omega : \varphi(x) \neq g(x)\} < \varepsilon$. See Rudin [37, p. 53].

We define a linear functional $\Lambda : X \to Y$ by $\Lambda g = kg$. We claim that Λ is continuous. By Theorem 3.1, we need only show that, for any $h \in L^\infty(\Omega)$, the map

$$f \mapsto \int_\Omega h(x)\,k(x)\,f(x)\,dx$$

is continuous on X. This follows from the fact that $hk \in L^1(\Omega)$, so that the map in question is an evaluation of the type that is rendered continuous by the topology $\sigma(L^\infty(\Omega), L^1(\Omega))$.

Then K is the image under the continuous map Λ of the unit ball in $L^\infty(\Omega)$, which is compact in X by Theorem 3.15. Thus K is compact in Y.

Now let f_i be a sequence in K; then $f_i = kg_i$, where g_i lies in the unit ball of $L^\infty(\Omega)$. (One may take $g_i(x) = f_i(x)/k(x)$ when $k(x) \neq 0$, and $g_i(x) = 0$ otherwise.) Because $L^1(\Omega)$ is separable, the weak* topology on the ball is metrizable (Theorem 3.21). Thus, a subsequence g_{i_j} converges weak* to a limit g in $L^\infty(\Omega)$. This means that

$$\int_\Omega g_{i_j}(x)\,h(x)\,dx \to \int_\Omega g(x)\,h(x)\,dx \quad \forall\, h \in L^1(\Omega).$$

It follows that

$$\int_\Omega g_{i_j}(x)\,k(x)\,u(x)\,dx \to \int_\Omega g(x)\,k(x)\,u(x)\,dx \quad \forall\, u \in L^\infty(\Omega),$$

which implies that $f_{i_j} := kg_{i_j}$ converges weakly in $L^1(\Omega)$ to gk. □

6.18 Exercise. For each $x \in \Omega$, let $F(x)$ be a closed convex subset of $\mathbb{R}$ satisfying $|F(x)| \leqslant k(x)$. Prove that the set

$$\Phi = \big\{ f \in L^1(\Omega) : f(x) \in F(x),\; x \in \Omega \text{ a.e.} \big\}$$

is sequentially weakly compact in $L^1(\Omega)$. □

6.19 Exercise. A *sawtooth function* x on $[0,1]$ is a Lipschitz, piecewise affine function $x : [0,1] \to \mathbb{R}$ with $x(0) = x(1) = 0$ such that $|x'(t)| = 1$ a.e. Let x_i be a sequence of such functions satisfying

$$\|x_i\|_{C[0,1]} \leqslant 1/i,$$

and set $v_i = x_i'$. Prove that v_i converges weakly in $L^1(0,1)$ to 0. Deduce that the set

$$\big\{ f \in L^1(0,1) : f(x) \in \{-1,1\} \text{ a.e.} \big\}$$

is not weakly compact. □

6.2 Measurable multifunctions

Let Ω be a subset of $\mathbb{R}^m$. A *multifunction* Γ from Ω to $\mathbb{R}^n$ is a mapping from Ω to the subsets of $\mathbb{R}^n$; thus, we associate with each $x \in \Omega$ a set $\Gamma(x)$ in $\mathbb{R}^n$, possibly the empty set. Such mappings arise rather frequently later on, and a recurrent issue will be that of finding a *measurable selection* of Γ. This means a measurable function $\gamma : \Omega \to \mathbb{R}^n$ such that $\gamma(x)$ belongs to $\Gamma(x)$ for almost all $x \in \Omega$.

Consider the following simple example, in which $n = m$. Let U be an open convex subset of $\mathbb{R}^n$, and $f : U \to \mathbb{R}$ a convex function. We have learned that the subdifferential $\Gamma(x) := \partial f(x)$ is nonempty for each $x \in U$. It follows from the axiom of choice that there is a function $\zeta : U \to \mathbb{R}^n$ such that $\zeta(x) \in \partial f(x)\ \ \forall x \in U$. Is there, however, a *measurable* function having this property?

Answering a question such as this requires a theory. We develop one in this section, in the context of Euclidean spaces.

Notation. We write $\Gamma : \Omega \rightsquigarrow \mathbb{R}^n$ to denote a multifunction Γ that maps a subset Ω of $\mathbb{R}^m$ to the subsets of $\mathbb{R}^n$.

One of the major ingredients in the theory is the following extension to multifunctions of the concept of measurable function.

Measurable multifunctions. The multifunction $\Gamma : \Omega \rightsquigarrow \mathbb{R}^n$ is *measurable* provided that Ω is measurable, and provided that the set

$$\Gamma^{-1}(V) = \big\{ x \in \Omega : \Gamma(x) \cap V \neq \emptyset \big\}$$

is (Lebesgue) measurable for every closed subset V of $\mathbb{R}^n$.

We obtain an equivalent definition by taking compact sets V in the definition. To see this, observe that any closed set V is the union of countably many compact sets V_i. Then we have

$$\Gamma^{-1}(V) = \bigcup_{i \geqslant 1} \Gamma^{-1}(V_i)\,.$$

If each $\Gamma^{-1}(V_i)$ is measurable, then so is $\Gamma^{-1}(V)$, as the countable union of measurable sets. The reader may show by a somewhat similar argument that when Γ is measurable, then the set $\Gamma^{-1}(V)$ is measurable for every *open* set V (this property, however, does not characterize measurability).

6.20 Exercise. Suppose that Γ is a singleton $\{\gamma(x)\}$ for each x. Prove that the multifunction Γ is measurable if and only if the function γ is measurable. □

The *effective domain* $\operatorname{dom}\Gamma$ of the multifunction $\Gamma : \Omega \rightsquigarrow \mathbb{R}^n$ is defined as follows:

$$\operatorname{dom}\Gamma = \big\{ x \in \Omega : \Gamma(x) \neq \emptyset \big\}.$$

By taking $V = \mathbb{R}^n$ in the definition of measurability, it follows that the effective domain of a measurable multifunction is measurable. We remark that as in the case of a function, redefining Γ on a set of measure zero does not affect its measurability, so in discussing measurable multifunctions we deal implicitly with equivalence classes, as we do with Lebesgue spaces.

6.21 Exercise. Let $u : \mathbb{R}^m \to \mathbb{R}^n$ and $r : \mathbb{R}^m \to \mathbb{R}_+$ be measurable functions, and let W be a measurable subset of $\mathbb{R}^n$. Prove that the multifunction Γ from $\mathbb{R}^m$ to $\mathbb{R}^n$ defined by $\Gamma(x) = W + B(u(x), r(x))$ is measurable. □

It is not hard to show that if γ_i is a sequence of measurable functions, then the multifunction $\Gamma(x) = \{\, \gamma_i(x) : i \geqslant 1 \,\}$ is measurable. The following shows that all *closed-valued* measurable multifunctions are in fact generated this way. (Γ is said to be closed-valued, of course, when $\Gamma(x)$ is a closed set for each $x \in \Omega$.)

6.22 Theorem. *Let $\Gamma : \Omega \rightsquigarrow \mathbb{R}^n$ be closed-valued and measurable. Then there exists a countable family $\{\, \gamma_i : \operatorname{dom} \Gamma \to \mathbb{R}^n \,\}$ of measurable functions such that*

$$\Gamma(x) = \operatorname{cl}\big\{\, \gamma_i(x) : i \geqslant 1 \,\big\}, \quad x \in \operatorname{dom} \Gamma \ \text{a.e.}$$

Proof. Let $\Delta = \operatorname{dom} \Gamma$. We begin by noting that, for any u in $\mathbb{R}^n$, the function $s \to d_{\Gamma(s)}(u)$ restricted to Δ is measurable (where $d_{\Gamma(s)}$ is as usual the Euclidean distance function). This follows from the identity (for $0 \leqslant r < R$)

$$d_{\Gamma(\cdot)}(u)^{-1}(r, R) = \{ s \in \Delta : \Gamma(s) \cap B(u, r) = \emptyset \} \cap \Gamma^{-1}(B^\circ(u, R)).$$

Now let $\{\, u_j \,\}_{j \geqslant 1}$ be a dense sequence in $\mathbb{R}^n$, and define a function $f_0 \colon \Delta \to \mathbb{R}^n$ as follows:

$$f_0(s) = \ \text{the first } u_j \text{ such that } d_{\Gamma(s)}(u_j) \leqslant 1.$$

Lemma. *The functions $s \to f_0(s)$ and $s \to d_{\Gamma(s)}(f_0(s))$ are measurable on Δ.*

To see this, observe that f_0 assumes countably many values, and that we have, for each $i \geqslant 1$:

$$\big\{\, s \colon f_0(s) = u_i \,\big\} = \bigcap_{j=1}^{i-1} \big\{\, s \colon d_{\Gamma(s)}(u_j) > 1 \,\big\} \bigcap \big\{\, s \colon d_{\Gamma(s)}(u_i) \leqslant 1 \,\big\}.$$

This implies that f_0 is measurable. Since the function $(s, u) \mapsto d_{\Gamma(s)}(u)$ is measurable in s and continuous in u, it is known (and actually proved in the next section, in the midst of more general goings on) that the function $s \to d_{\Gamma(s)}(f_0(s))$ is measurable. The lemma is proved.

We pursue the process begun above by defining for each integer $i \geqslant 0$ a function f_{i+1} such that $f_{i+1}(s)$ is the first u_j for which *both* the following hold:

$$|u_j - f_i(s)| \leqslant \tfrac{2}{3}\, d_{\Gamma(s)}\big(f_i(s)\big), \quad d_{\Gamma(s)}(u_j) \leqslant \tfrac{2}{3}\, d_{\Gamma(s)}\big(f_i(s)\big).$$

It follows as above that each f_i is measurable. Moreover, we deduce

$$d_{\Gamma(s)}\big(f_{i+1}(s)\big) \leqslant \big(\tfrac{2}{3}\big)^{i+1} d_{\Gamma(s)}\big(f_0(s)\big) \leqslant \big(\tfrac{2}{3}\big)^{i+1},$$

together with $|f_{i+1}(s) - f_i(s)| \leqslant (2/3)^{i+1}$. This implies that $\{f_i(s)\}$ is a Cauchy sequence converging for each $s \in \Delta$ to a value which we denote by $\gamma_0(s)$, and that $\gamma_0(x) \in \Gamma(x)$ a.e. in Δ. As a limit of measurable functions, γ_0 is measurable.

For every pair of positive integers i, j, we define a multifunction $\Gamma_{i,j} : \Omega \rightsquigarrow \mathbb{R}^n$ as follows:

$$\Gamma_{i,j}(x) = \begin{cases} \emptyset & \text{if } x \notin \Delta \\ \Gamma(x) \cap B(u_i, 1/j) & \text{if } x \in \Delta \text{ and } \Gamma(x) \cap B(u_i, 1/j) \neq \emptyset \\ \{\gamma_0(x)\} & \text{otherwise.} \end{cases}$$

For any closed subset V of $\mathbb{R}^n$, the set $\Gamma_{i,j}^{-1}(V)$ is given by

$$\big\{x : \Gamma(x) \cap V \cap B(u_i, 1/j) \neq \emptyset\big\} \bigcup \Big[\big\{x \in \Delta : \Gamma(x) \cap B(u_i, 1/j) = \emptyset\big\} \cap \gamma_0^{-1}(V)\Big].$$

It follows that $\Gamma_{i,j}$ is measurable and closed-valued; its effective domain is Δ. By the argument above (applied to $\Gamma_{i,j}$ rather than Γ), there exists a measurable function $\gamma_{i,j}$ such that $\gamma_{i,j}(x) \in \Gamma_{i,j}(x)$, $x \in \Delta$ a.e.

We claim that the countable collection $\gamma_{i,j}$, together with γ_0, satisfies the conclusion of the theorem.

To see this, let $S_{i,j}$ be the null set of $x \in \Delta$ for which the inclusion $\gamma_{i,j}(x) \in \Gamma(x)$ fails. Now let $x \in \Delta \setminus [\cup_{i,j} S_{i,j}]$, and fix any $\gamma \in \Gamma(x)$, $\gamma \neq \gamma_0(x)$. There exists a sequence u_{i_k} in $\{u_i\}$ and an increasing sequence of integers $j_k \to \infty$ such that $|u_{i_k} - \gamma| < 1/j_k$. Then we have

$$\gamma_{i_k, j_k}(x) \in B(u_{i_k}, 1/j_k) \implies |\gamma_{i_k, j_k}(x) - u_{i_k}| < 1/j_k \implies |\gamma_{i_k, j_k}(x) - \gamma| < 2/j_k.$$

Thus, $\Gamma(x) = \operatorname{cl}\{\gamma_{i,j}(x)\}$, $x \in \Delta$ a.e. □

6.23 Corollary. (Measurable selections) *Let $\Gamma : \Omega \rightsquigarrow \mathbb{R}^n$ be closed-valued and measurable. Then there exists a measurable function $\gamma : \operatorname{dom}\Gamma \to \mathbb{R}^n$ such that*

$$\gamma(x) \in \Gamma(x), \ x \in \operatorname{dom}\Gamma \text{ a.e.}$$

6.24 Exercise. Let $\Gamma : \Omega \rightsquigarrow \mathbb{R}^n$ and $G : \Omega \rightsquigarrow \mathbb{R}^n$ be two measurable closed-valued multifunctions. Prove that $\Gamma + G$ is measurable. □

The measurable multifunctions that the reader is likely to encounter will most often have the following structure.

6.25 Proposition. *Let Ω be a measurable subset of $\mathbb{R}^m$, and $\varphi : \Omega \times \mathbb{R}^n \times \mathbb{R}^\ell \to \mathbb{R}$ a function with the following properties:*

- *The mapping $x \mapsto \varphi(x,p,q)$ is measurable on Ω for each $(p,q) \in \mathbb{R}^n \times \mathbb{R}^\ell$, and*
- *The mapping $(p,q) \mapsto \varphi(x,p,q)$ is continuous for each $x \in \Omega$.*

Let $P, Q : \Omega \rightsquigarrow \mathbb{R}^n$ be measurable closed-valued multifunctions, and $c,d : \Omega \to \mathbb{R}$ measurable functions. Then $\Gamma : \Omega \rightsquigarrow \mathbb{R}^n$ defined by

$$\Gamma(x) = \big\{ p \in P(x) : c(x) \leqslant \varphi(x,p,q) \leqslant d(x) \text{ for some } q \in Q(x) \big\}$$

is measurable.

Proof. Let p_i be a countable family of measurable selections of P that generate the multifunction P as described in Theorem 6.22, and similarly, let q_i generate Q. Let Δ_P and Δ_Q be the effective domains of P and Q.

Then, if V is a compact subset of $\mathbb{R}^n$, it follows (fairly easily, though we beg the reader's indulgence for the next expression) that

$$\Gamma^{-1}(V) = \bigcup_{i \geqslant 1} \bigcap_{j \geqslant 1} \bigcup_{k \geqslant 1} \Big\{ x \in \Delta_P \cap \Delta_Q :$$
$$p_k(x) \in \big(V + j^{-1}B\big),\ |q_k(x)| \leqslant i,\ c(x) - j^{-1} < \varphi\big(x, p_k(x), q_k(x)\big) < d(x) + j^{-1} \Big\}.$$

This is recognized to be a measurable set, since the function

$$x \mapsto \varphi(x, p_k(x), q_k(x))$$

is measurable (a known result on measurable functions, see Props. 6.34 and 6.35 below). □

6.26 Corollary. *The intersection of two closed-valued measurable multifunctions $\Gamma_1, \Gamma_2 : \Omega \rightsquigarrow \mathbb{R}^n$ is measurable.*

Proof. Let Δ_1 and Δ_2 be the effective domains of Γ_1 and Γ_2. Define a function φ on $\Omega \times \mathbb{R}^n$ by

$$\varphi(x,p) = \begin{cases} d_{\Gamma_1(x)}(p) + d_{\Gamma_2(x)}(p) & \text{if } x \in \Delta_1 \cap \Delta_2, \\ -1 & \text{otherwise.} \end{cases}$$

The proof of Theorem 6.22 showed that φ is measurable in x; it is evidently continuous in p. Then the multifunction

$$\Gamma_1(x) \cap \Gamma_2(x) = \big\{ p : \varphi(x,p) = 0 \big\}$$

is measurable by Prop. 6.25. □

The **graph** of a multifunction $\Gamma : \Omega \rightsquigarrow \mathbb{R}^n$ is the set

$$\operatorname{gr}\Gamma = \{ (x,\gamma) \in \Omega \times \mathbb{R}^n : \gamma \in \Gamma(x) \}.$$

6.27 Corollary. *Let $\Omega \subset \mathbb{R}^m$ be measurable. If $\Gamma : \Omega \rightsquigarrow \mathbb{R}^n$ has the property that $\operatorname{gr}\Gamma$ is closed, then Γ is measurable.*

Proof. We may assume that $\operatorname{gr}\Gamma \neq \emptyset$; then the function $(x,v) \mapsto d_{\operatorname{gr}\Gamma}(x,v)$ is continuous. For any $x \in \Omega$, the set $\Gamma(x)$ is given by $\{ v \in \mathbb{R}^n : d_{\operatorname{gr}\Gamma}(x,v) = 0 \}$, which leads to the required conclusion with the help of Prop. 6.25. □

6.28 Corollary. *Let $G : \Omega \rightsquigarrow \mathbb{R}^n$ be measurable and closed-valued. Then the multifunction Γ defined by $\Gamma(x) = \operatorname{co} G(x)$ is measurable.*

Proof. Let Σ denote the set of all nonnegative vectors $\lambda = (\lambda_0, \lambda_1, \dots, \lambda_n) \in \mathbb{R}^{n+1}$ whose coordinates sum to 1. It is not hard to see that the multifunction

$$Q(x) = \Sigma \times G(x) \times G(x) \times \cdots \times G(x)$$

is measurable (where the Cartesian product contains $n+1$ factors equal to $G(x)$). Let f be defined by

$$f(\lambda, g_0, g_1, \dots, g_n) = \sum_{i=0}^n \lambda_i g_i,$$

where each g_i lies in $\mathbb{R}^n$. Then, by Prop. 2.6, the set $\Gamma(x)$ is described by

$$\{ v \in \mathbb{R}^n : |v - f(\lambda, g_0, g_1, \dots, g_n)| = 0 \text{ for some } (\lambda, g_0, g_1, \dots, g_n) \in Q(x) \}.$$

The result now follows from Prop. 6.25. □

6.29 Proposition. *Let $\Omega \subset \mathbb{R}^m$ be measurable, and let $G : \Omega \rightsquigarrow \mathbb{R}^n$ be a multifunction whose values are nonempty compact convex sets. Let $H_{G(x)}(\cdot)$ be the support function of the set $G(x)$. Then G is measurable if and only if, for any $p \in \mathbb{R}^n$, the function $x \mapsto H_{G(x)}(p)$ is measurable on Ω.*

Proof. Suppose first that the support function has the stated measurability property. Let V be a nonempty compact subset of $\mathbb{R}^n$, and let $\{ v_j \}$ be a countable dense set in V. Let $\{ p_k \}$ be a countable dense set in $\mathbb{R}^n$. Then (invoking the separation theorem for the last step) it follows that

$$\begin{aligned} \{ x \in \Omega : G(x) \cap V = \emptyset \} &= \bigcup_{\varepsilon > 0} \{ x \in \Omega : G(x) \cap [V + \varepsilon B] = \emptyset \} \\ &= \bigcup_{i \geqslant 1} \bigcap_{j \geqslant 1} \{ x \in \Omega : G(x) \cap B(v_j, i^{-1}) = \emptyset \} = \\ &\bigcup_{i \geqslant 1} \bigcap_{j \geqslant 1} \bigcup_{k \geqslant 1} \{ x \in \Omega : H_{G(x)}(p_k) < \langle p_k, v_j \rangle + i^{-1} |p_k| \}. \end{aligned}$$

This implies that $\{ x \in \Omega : G(x) \cap V = \emptyset \}$ is measurable, so G is measurable.

Conversely, let G be measurable, and let the functions γ_i generate G as in Theorem 6.22. Then we have

$$H_{G(x)}(p) = \sup \left\{ \langle p, \gamma_i(x) \rangle : i \geqslant 1 \right\} \ \forall x \in \Omega,$$

which reveals the required measurability in x of the function on the left. □

6.30 Exercise. Let $G : \Omega \rightsquigarrow \mathbb{R}^n$ be measurable and closed-valued. If $u : \mathbb{R}^m \to \mathbb{R}^n$ is measurable, prove that the function $x \mapsto d_{G(x)}(u(x))$ is measurable on $\operatorname{dom} G$. □

The following fact, already invoked in proving Theorem 6.10, will be useful again later. It bears upon interchanging the integral and the supremum.

6.31 Theorem. *Let Ω be an open subset of $\mathbb{R}^m$, and let $\varphi : \Omega \times \mathbb{R}^n \to \mathbb{R}$ be a function such that $\varphi(x,p)$ is measurable in the x variable and continuous in the p variable. Let $P : \Omega \rightsquigarrow \mathbb{R}^n$ be measurable and closed-valued. Let Σ denote the set of all functions $p \in L^\infty(\Omega, \mathbb{R}^n)$ which satisfy*

$$p(x) \in P(x), \ \ x \in \Omega \text{ a.e.},$$

and for which the integral

$$\int_\Omega \varphi\big(x, p(x)\big)\, dx$$

is well defined, either finitely or as $+\infty$. Then, if Σ is nonempty, the integral

$$\int_\Omega \sup_{p \in P(x)} \varphi(x,p)\, dx$$

is well defined, either finitely or as $+\infty$, and we have

$$\int_\Omega \sup_{p \in P(x)} \varphi(x,p)\, dx = \sup_{p(\cdot) \in \Sigma} \int_\Omega \varphi\big(x, p(x)\big)\, dx,$$

where both sides may equal $+\infty$.

Proof. The hypotheses imply that $P(x) \neq \emptyset$ for $x \in \Omega$ a.e. By Theorem 6.22, there exists a countable collection $\{ p_i \}$ of measurable selections of P such that

$$P(x) = \operatorname{cl}\{ p_i(x) \}, \ \ x \in \Omega \text{ a.e.}$$

Since $\varphi(x, \cdot)$ is continuous, we have

$$\sigma(x) := \sup_{p \in P(x)} \varphi(x,p) = \sup_{i \geqslant 1} \varphi\big(x, p_i(x)\big) \text{ a.e.},$$

which shows that $\sigma : \Omega \to \mathbb{R}_\infty$ is measurable, as a countable supremum of measurable functions.

Let $\bar{p}$ be any element of Σ. Then $\sigma(x)$ is bounded below by the function $\varphi(x, \bar{p}(x))$, and it follows that the integral of σ is well defined, possibly as $+\infty$; this is the first assertion of the theorem.[4]

If the integral over Ω of the function $\varphi(x, \bar{p}(x))$ is $+\infty$, then the remaining assertion is evident. We may proceed under the assumption, therefore, that the integral in question is finite. Fix a positive integer N, and define

$$\sigma_N(x) = \sup \left\{ \varphi(x,p) : p \in P(x) \cap B(\bar{p}(x), N) \right\}.$$

Using Cor. 6.26, and arguing as above, we find that σ_N is measurable. Evidently we have $\varphi(x,\bar{p}(x)) \leqslant \sigma_N(x) \leqslant \sigma(x)$, $x \in \Omega$ a.e. The multifunction

$$\Gamma(x) = \left\{ p \in P(x) \cap B(\bar{p}(x), N) : \sigma_N(x) = \varphi(x,p) \right\}$$

is measurable by Prop. 6.25; since its values on Ω are closed and nonempty, it admits a measurable selection p_N. It follows that $p_N \in \Sigma$, whence

$$\sup_{p(\cdot) \in \Sigma} \int_\Omega \varphi\big(x, p(x)\big)\, dx \geqslant \int_\Omega \varphi\big(x, p_N(x)\big)\, dx \to \int_\Omega \sigma(x)\, dx,$$

by monotone convergence. But the supremum on the left in this expression is evidently bounded above by the integral on the right (and neither depend on N). Thus we obtain equality. ☐

The point of the next result is that a local minimum in $L^1(\Omega)$ translates into a global minimum (almost everywhere) at the pointwise level.

6.32 Theorem. *Let Ω, φ, P, and Σ be as described in* Theorem 6.31, *and let $\bar{p} \in \Sigma$ be such that the integral*

$$\int_\Omega \varphi(x, \bar{p}(x))\, dx$$

is finite. Suppose that for some $\delta > 0$, we have:

$$p(\cdot) \in \Sigma, \int_\Omega |p(x) - \bar{p}(x)|\, dx \leqslant \delta \implies \int_\Omega \varphi\big(x, p(x)\big)\, dx \geqslant \int_\Omega \varphi\big(x, \bar{p}(x)\big).$$

Then, for almost every $x \in \Omega$, we have $\varphi(x,p) \geqslant \varphi(x,\bar{p}(x))\ \forall\, p \in P(x)$.

Proof. We reason by the absurd. If the conclusion fails, there exist positive numbers ε and M such that the multifunction

$$\Gamma(x) = \left\{ p \in P(x) \cap B(\bar{p}(x), M) : \varphi\big(x, \bar{p}(x)\big) - M \leqslant \varphi(x,p) \leqslant \varphi\big(x, \bar{p}(x)\big) - \varepsilon \right\}$$

---

[4] We have used the following fact from integration: if two measurable functions f and g satisfy $f \geqslant g$, and if the integral of g is well defined, either finitely or as $+\infty$, then the integral of f is well defined, either finitely or as $+\infty$.

has an effective domain of positive measure. Then, for any $m > 0$ sufficiently small, we may use a measurable selection γ of Γ to define a function p as follows: let S_m be a measurable subset of $\operatorname{dom}\Gamma$ satisfying $\operatorname{meas}(S_m) = m$, and set $p(x) = \gamma(x)$ if $x \in S_m$, and $p(x) = \bar{p}(x)$ otherwise. It follows that $p \in \Sigma$. But for m sufficiently small, p satisfies

$$\int_\Omega |p(x) - \bar{p}(x)|\,dx \leqslant \delta\,, \quad \int_\Omega \varphi\big(x, p(x)\big)\,dx < \int_\Omega \varphi\big(x, \bar{p}(x)\big)\,,$$

which is the desired contradiction. □

6.3 Integral functionals and semicontinuity

A technical issue of some importance to us later concerns the measurability of certain composite functions f arising in the following way:

$$f(t) = \Lambda\big(t, x(t), x'(t)\big)\,.$$

Here, x is an element of $\mathrm{AC}[0,1]$ (say), so that $x'(\cdot)$ is merely Lebesgue measurable. When the function $\Lambda(t,x,v)$ is continuous (in all its variables (t,x,v)), then, as the reader will recall, it is a basic result in measurability theory that f is Lebesgue measurable (a continuous function of a measurable function is measurable). This is a minimal requirement for considering the integral of f, as we do later in the calculus of variations. When Λ is less than continuous, the issue is more complex.

To illustrate this point, let S be a non measurable subset of $[0,1]$, and define a subset G of $\mathbb{R}^2$ as follows:

$$G = \big\{ (s,s) : s \in S \big\}.$$

Since G is a subset of the diagonal in $\mathbb{R}^2$, a set of measure zero, it follows that G is a null set for two-dimensional Lebesgue measure, which is complete. Thus G is a measurable set and its characteristic function χ_G is measurable.

Let us define $\Lambda(t,x,v) = \Lambda(t,x) = 1 - \chi_G(t,x)$, a measurable function. The reader may verify that $\Lambda(t,x)$ is lower semicontinuous separately in each variable; in particular, measurable as a function of t for each x, and lower semicontinuous as a function of x for each t. When we proceed to substitute the function $x(t) = t$ into Λ, we obtain

$$f(t) = \Lambda(t,t) = 1 - \chi_G(t,t).$$

This function f *fails* to be measurable, however, since

$$\big\{ t \in [0,1] : f(t) < 1/2 \big\} = \big\{ t \in [0,1] : \chi_G(t,t) = 1 \big\} = S$$

is not a measurable set.

LB measurability. We shall prove below that when Λ is measurable in t and continuous in (x,v), then the composition f is measurable, as desired. The example above demonstrates, however, that when Λ fails to be continuous in (x,v), as it will later on occasion, mere measurability in (x,v), or even lower semicontinuity (which is a natural hypothesis in the contexts to come), does not suffice to guarantee the measurability of f.

One could compensate for the lack of continuity by simple requiring that Λ, as a function of (t,x,v), be Borel measurable. This is because of the fact that the composition of a Borel measurable function with a measurable one is measurable. Since lower semicontinuous functions are Borel measurable, it follows, as a special case, that our measurability concerns would disappear if we took Λ to be lower semicontinuous in the entirety of its variables (t,x,v). This is overly restrictive as a global hypothesis, however, and even Borel measurability is asking too much, since mere Lebesgue measurability in t is desirable in certain applications.

A more common way to deal with the measurability issue is to employ a hybrid hypothesis of the following type:

6.33 Definition. *A function $F : \mathbb{R}^m \times \mathbb{R}^n \to \mathbb{R}$ of two variables (x,y), where $x \in \mathbb{R}^m$ and $y \in \mathbb{R}^n$, is said to be* LB *measurable in x and y when it is the case that F is measurable with respect to the σ-algebra $L \times B$ generated by products of Lebesgue measurable subsets of $\mathbb{R}^m$ (for x) and Borel measurable subsets of $\mathbb{R}^n$ (for y).*

Can the reader can go so far back as to remember that a σ-algebra is a family of sets closed under taking complements, countable intersections, and countable unions? We remark that if $F : \mathbb{R}^m \times \mathbb{R}^n \to \mathbb{R}$ is lower semicontinuous, then F is Borel measurable, which implies that F is LB measurable.

Returning to the context of the function Λ, we say that $\Lambda(t,x,v)$ is LB measurable if Λ is LB measurable in the variables t and (x,v); that is, measurable with respect to the σ-algebra generated by products of Lebesgue measurable subsets of $[a,b]$ (for t) and Borel measurable subsets of $\mathbb{R}^2$ (for (x,v)). This property guarantees the measurability of the function f above, and is satisfied in the important case in which $\Lambda(t,x,v)$ is measurable with respect to t and continuous with respect to (x,v), as we proceed to show in the next two results.

6.34 Proposition. *Let F be* LB *measurable as in* Def. 6.33, *and let $g : \mathbb{R}^m \to \mathbb{R}^n$ be Lebesgue measurable. Then the mapping $x \mapsto F(x, g(x))$ is Lebesgue measurable.*

Proof. Let U be a Lebesgue measurable set in $\mathbb{R}^m$ and V a Borel set in $\mathbb{R}^n$. Then the set

$$\left\{ x \in \mathbb{R}^m : \big(x, g(x)\big) \in U \times V \right\} = U \cap g^{-1}(V)$$

is clearly Lebesgue measurable. Let us denote by A the collection of all subsets S of $\mathbb{R}^m \times \mathbb{R}^n$ having the property that the set

$$\{ x \in \mathbb{R}^m : (x, g(x)) \in S \}$$

is Lebesgue measurable. It is easy to verify that A is a σ-algebra, and it contains the products $U \times V$. It follows that A contains the σ-algebra $L \times B$ generated by products of Lebesgue measurable subsets of $\mathbb{R}^m$ and Borel measurable subsets of $\mathbb{R}^n$.

Now let W be any open subset of $\mathbb{R}^n$. Since F is LB measurable, the set $F^{-1}(W)$ is LB measurable, and hence lies in A. As a consequence of this fact, we deduce that the set

$$\{ x : F\big(x, g(x)\big) \in W \} = \{ x : \big(x, g(x)\big) \in F^{-1}(W) \}$$

is Lebesgue measurable. This confirms the required measurability. □

6.35 Proposition. *If a function $F : \mathbb{R}^m \times \mathbb{R}^n \to \mathbb{R}$ of two variables (x, y) is measurable in x and continuous in y, then F is* LB *measurable in x and y.*

Proof. Let $\{u_i\}$ be a dense sequence in $\mathbb{R}^n$, and for each positive integer k define a function $f_k(x, y) = F(x, u_j)$, where u_j is the first point of the dense set satisfying $|u_j - y| \leqslant k^{-1}$. (Thus, j depends on y and k.) Then $F(x, y) = \lim_{k\to\infty} f_k(x, y)$ for every (x, y), by the continuity of F in y.

It suffices therefore to prove that each f_k is LB measurable. Let W be any open subset of $\mathbb{R}$. Then the set $f_k^{-1}(W)$ is the union over $j \geqslant 1$ of the sets

$$\{ x : F(x, u_j) \in W \} \times \{ y : |u_j - y| \leqslant k^{-1} \text{ and } |u_i - y| > k^{-1} \ (i = 1, \dots, j-1) \}.$$

This reveals $f_k^{-1}(W)$ to be a countable union of products of the type which generate the σ-algebra $L \times B$, whence the required measurability. □

We remark that a function F of two variables having the properties described in Prop. 6.35 is often referred to as a **Carathéodory function**.

The next result says that inserting a measurable function into a continuous slot preserves LB measurability.

6.36 Proposition. *Let $F : \mathbb{R}^m \times \mathbb{R}^\ell \times \mathbb{R}^n \to \mathbb{R}$ satisfy the following:*

(a) *The map $(x, z) \mapsto F(x, u, z)$ is* LB *measurable for each u;*

(b) *The function $u \mapsto F(x, u, z)$ is continuous for each (x, z).*

Then, for any Lebesgue measurable function $g : \mathbb{R}^m \to \mathbb{R}^\ell$, the function

$$(x, z) \mapsto F(x, g(x), z)$$

is LB *measurable.*

Proof. Let $\{u_i\}$ be a dense sequence in $\mathbb{R}^\ell$, and for each positive integer k define a function $f_k(x, z) = F(x, u_j, z)$, where u_j is the first term of the sequence satisfying

$|u_j - g(x)| \leqslant k^{-1}$. (Thus, j depends on x and k.) Then $f_k(x,z)$ converges pointwise to $F(x, g(x), z)$, so it suffices to prove that each f_k is LB measurable. This follows from the identity

$$f_k^{-1}(W) = \bigcup_{j \geqslant 1} \{ (x,z) : F(x,u_j,z)) \in W \} \cap \\ \{ (x,z) : |u_j - g(x)| \leqslant k^{-1} \text{ and } |u_i - g(x)| > k^{-1} \, (i = 1, \dots, j-1) \}$$

(where W is any open subset of $\mathbb{R}$), which expresses $f_k^{-1}(W)$ as a countable union of sets belonging to $L \times B$. □

Semicontinuity of integral functionals. Let Ω be an open subset of $\mathbb{R}^m$. We study the semicontinuity of the following integral functional:

$$J(u,z) = \int_\Omega F\big(x, u(x), z(x)\big)\, dx.$$

Here, $F : \Omega \times \mathbb{R}^\ell \times \mathbb{R}^n \to \mathbb{R}$ is a function whose three arguments are generically denoted by x, u, z. We are also given a subset Q of $\Omega \times \mathbb{R}^\ell$ which defines a restriction on the functions $u : \Omega \to \mathbb{R}^\ell$ involved in the discussion: they must satisfy

$$\big(x, u(x)\big) \in Q, \;\; x \in \Omega \text{ a.e.}$$

We shall impose the following hypotheses on the data:

6.37 Hypothesis.

(a) *F is lower semicontinuous in (u,z), and convex with respect to z;*

(b) *For every measurable $u : \Omega \to \mathbb{R}^\ell$ having $(x, u(x)) \in Q$ a.e., and for every measurable $z : \Omega \to \mathbb{R}^n$, the function $x \mapsto F(x, u(x), z(x))$ is measurable;*

(c) *There exist $\alpha \in L^1(\Omega)$, $\beta \in L^\infty(\Omega, \mathbb{R}^n)$ such that*

$$F(x,u,z) \geqslant \alpha(x) + \langle \beta(x), z \rangle \;\; \forall (x,u) \in Q, \; z \in \mathbb{R}^n;$$

(d) *Q is closed in $\Omega \times \mathbb{R}^\ell$.*

An immediate consequence of these hypotheses is that, for every measurable function $u : \Omega \to \mathbb{R}^\ell$ having $(x, u(x)) \in Q$ a.e., and for every summable function $z : \Omega \to \mathbb{R}^n$, the function

$$x \mapsto F(x, u(x), z(x))$$

is measurable, and it is bounded below as follows:

$$F(x, u(x), z(x)) \geqslant \alpha(x) + \langle \beta(x), z(x) \rangle, \;\; x \in \Omega \text{ a.e.}$$

Since the right side of this inequality is summable over Ω, the integral $J(u,z)$ is well defined, either finitely or as $+\infty$ (a fact from integration theory). It is the lower semicontinuity of J that is the essential point being considered.

Remark. In view of our earlier results, we are able to identify certain cases in which hypothesis (b) above is guaranteed to hold:

- $F(x,u,z)$ is measurable in x and continuous in (u,z) (see Prop. 6.35);
- $F(x,u,z)$ is LB measurable in x and (u,z) (by Prop. 6.34);
- $F(x,u,z)$ is continuous in u and LB measurable in (x,z) (by Prop. 6.36).

The theorem below will provide one of the main ingredients in the recipe that we call the direct method.

6.38 Theorem. (Integral semicontinuity) *Let u_i be a sequence of measurable functions on Ω having $(x,u_i(x)) \in Q$ a.e. which converges almost everywhere to a limit u_*. Let z_i be a sequence of functions converging weakly in $L^r(\Omega,\mathbb{R}^n)$ to z_*, where $r>1$. Then*

$$J(u_*,z_*) \leqslant \liminf_{i\to\infty} J(u_i,z_i).$$

Proof. Fix $\delta>0$, and define, for $(x,u)\in Q,\ z\in\mathbb{R}^n$, the function

$$H(x,u,p) = \sup\left\{ \langle p,w\rangle - F(x,u,w) - \delta|w|^r/r : w\in\mathbb{R}^n \right\}.$$

The properties of H play an essential role in the proof.

Lemma 1. *There is a positive constant c such that*

$$H(x,u,p) \leqslant c|p-\beta(x)|^{r_*} - \alpha(x)\ \ \forall (x,u)\in Q,\ p\in\mathbb{R}^m,$$

where r_* is the conjugate exponent to r.

Proof. Observe that, by the inequality in Hypothesis 6.37 (c), we have:

$$\begin{aligned} H(x,u,p) &= \sup_w \left\{ \langle p,w\rangle - F(x,u,w) - \delta|w|^r/r \right\} \\ &\leqslant \sup_w \left\{ \langle p,w\rangle - \alpha(x) - \langle \beta(x),w\rangle - \delta|w|^r/r \right\} \\ &= c|p-\beta(x)|^{r_*} - \alpha(x)\ \ \text{(by explicit calculation)} \end{aligned}$$

where $c := (r_*\delta^{r_*-1})^{-1}$. □

Lemma 2. *Fix $(x,u)\in Q$ and $p\in\mathbb{R}^n$. Then:*

(a) *The function $H(x,u,\cdot)$ is continuous at p.*

(b) *The function $v\mapsto H(x,v,p)$ is upper semicontinuous at u in the following sense:*

$$(x,v_i) \in Q \;\; \forall\, i, \;\; v_i \to u \implies H(x,u,p) \geqslant \limsup_{i\to\infty} H(x,v_i,p).$$

(c) *For all* $w \in \mathbb{R}^n$, *we have*

$$F(x,u,w) + \delta\,|\,w\,|^r/r = \sup_{p \in \mathbb{R}^n} \big\{ \langle\, w, p\,\rangle - H(x,u,p) \big\}.$$

Proof. Since the function $p \mapsto H(x,u,p)$ is convex and finite on $\mathbb{R}^n$ (by Lemma 1), we know it to be continuous, as affirmed in (a). We now fix x and p and turn to assertion (b).

Let v_i be a sequence converging to u for which $\lim_{i\to\infty} H(x,v_i,p) \geqslant \ell \in \mathbb{R}$. We establish (b) by proving that $H(x,u,p) \geqslant \ell$. Note that the supremum defining $H(x,v_i,p)$ may be restricted to those w satisfying

$$\langle\, p, w\,\rangle - \alpha(x) - \langle\, \beta(x), w\,\rangle - \delta\,|\,w\,|^r/r \;\geqslant\; H(x,v_i,p) - 1,$$

and consequently, to the points w in a compact set. It follows that the supremum is attained at a point w_i, and that the sequence w_i is bounded. Taking a subsequence if necessary, and without relabeling, we may assume $w_i \to w$. Invoking the lower semicontinuity of F, we have

$$\begin{aligned} H(x,u,p) &\geqslant \langle\, p, w\,\rangle - F(x,u,w) - \delta\,|\,w\,|^r/r \\ &\geqslant \langle\, p, w\,\rangle - \liminf_{i\to\infty} F\big(x,v_i,w_i\big) - \delta\,|\,w\,|^r/r \\ &= \limsup_{i\to\infty} \big\{ \langle\, p, w_i\,\rangle - F\big(x,v_i,w_i\big) - \delta\,|\,w_i\,|^r/r \big\} = \lim_{i\to\infty} H(x,v_i,p) \geqslant \ell, \end{aligned}$$

as required.

For the final assertion, note that $H(x,u,\cdot)$ is defined as the conjugate of the convex lower semicontinuous function

$$w \mapsto F(x,u,w) + \delta\,|\,w\,|^r/r.$$

Thus the equality is a consequence of Theorem 4.21. □

Lemma 3. *Let* $u : \Omega \to \mathbb{R}^\ell$ *be a measurable function having* $(x,u(x)) \in Q$ a.e., *and let* $p : \Omega \to \mathbb{R}^n$ *be measurable. Then the function* $x \mapsto H(x,u(x),p(x))$ *is measurable.*

Proof. Note that the function $w \mapsto F(x,u,w)$ is continuous, since it is convex and finite. It follows that if $\{\, w_i \,\}$ is a countable dense set in $\mathbb{R}^n$, we have (almost everywhere)

$$H\big(x,u(x),p(x)\big) = \sup_{i \geqslant 1} \big\{ \langle\, p(x), w_i\,\rangle - F(x,u(x),w_i) - \delta\,|\,w_i\,|^r/r \big\}.$$

Thus the left side is the upper envelope of a countable family of measurable functions, and is therefore measurable. □

Note that the limit function u_* satisfies

$$(x, u_*(x)) \in Q \text{ a.e.},$$

in view of Hypothesis 6.37 (d). We now write, without claiming that $J(u_*, z_*)$ is finite:

$$\begin{aligned}
J(u_*, z_*) &= \int_\Omega F\big(x, u_*(x), z_*(x)\big)\,dx \\
&\leqslant \int_\Omega \big\{ F\big(x, u_*(x), z_*(x)\big) + \delta\,|z_*(x)|^r \big\}\,dx \\
&= \int_\Omega \sup_{p\in\mathbb{R}^n} \big\{ \langle z_*(x), p\rangle - H(x, u_*(x), p) \big\}\,dx \ \text{(by (c) of Lemma 2)} \\
&= \sup_{p(\cdot)\in L^\infty(\Omega)} \int_\Omega \big\{ \langle z_*(x), p(x)\rangle - H\big(x, u_*(x), p(x)\big) \big\}\,dx
\end{aligned}$$

(we use Theorem 6.31 and Lemma 3 to switch integral and supremum)

$$\leqslant \sup_{p(\cdot)\in L^\infty(\Omega)} \left[\lim_{i\to\infty} \int_\Omega \langle p(x), z_i(x)\rangle\,dx - \int_\Omega \limsup_{i\to\infty} H\big(x, u_i(x), p(x)\big)\,dx \right]$$

(since z_i converges weakly to z_*, and u_i to u a.e., and since H is upper semicontinuous in u, by Lemma 2)

$$\leqslant \sup_{p(\cdot)\in L^\infty(\Omega)} \left[\lim_{i\to\infty} \int_\Omega \langle p(x), z_i(x)\rangle\,dx - \limsup_{i\to\infty} \int_\Omega H\big(x, u_i(x), p(x)\big)\,dx \right]$$

(Fatou 's lemma applies, since u_i has values in Q, $p \in L^\infty(\Omega)$, and the terms in H are uniformly integrably bounded above, by Lemma 1; Lemma 3 is used to assert that the integrand is measurable)

$$\begin{aligned}
&= \sup_{p(\cdot)\in L^\infty(\Omega)} \left[\liminf_{i\to\infty} \int_\Omega \big\{ \langle p(x), z_i(x)\rangle - H\big(x, u_i(x), p(x)\big) \big\}\,dx \right] \\
&\leqslant \liminf_{i\to\infty} \sup_{p(\cdot)\in L^\infty(\Omega)} \left[\int_\Omega \big\{ \langle p(x), z_i(x)\rangle - H\big(x, u_i(x), p(x)\big) \big\}\,dx \right] \\
&= \liminf_{i\to\infty} \int_\Omega \sup_{p\in\mathbb{R}^n} \big\{ \langle p, z_i(x)\rangle - H\big(x, u_i(x), p\big) \big\}\,dx \ \text{(by Theorem 6.31)} \\
&= \liminf_{i\to\infty} \int_\Omega \big\{ F\big(x, u_i(x), z_i(x)\big) + \delta\,|z_i(x)|^r \big\}\,dx \ \text{(by (c) of Lemma 2)} \\
&\leqslant \liminf_{i\to\infty} J(u_i, z_i) + \delta \limsup_{i\to\infty} \|z_i\|_r^r .
\end{aligned}$$

Since z_i is weakly convergent in $L^r(\Omega, \mathbb{R}^n)$, the sequence z_i is norm bounded, so that $\limsup_{i\to\infty} \| z_i \|_r^r$ is finite. Since $\delta > 0$ is arbitrary, we obtain the required conclusion.

We remark that the crux of the proof is to find a way to exploit the (merely) weak convergence of the sequence z_i; this has been done by rewriting certain expressions so as to have z_i appear only in *linear* terms. □

6.4 Weak sequential closures

In things to come, the reader will find that the closure properties of *differential inclusions* of the type

$$x'(t) \in \Gamma\big(t, x(t)\big),$$

where Γ is a multifunction, will play an important role. The following abstract result is a basic tool in this connection. Note that weak convergence in L^1 is now involved.

6.39 Theorem. (Weak closure) *Let $[a,b]$ be an interval in $\mathbb{R}$ and Q a closed subset of $[a,b] \times \mathbb{R}^\ell$. Let $\Gamma(t,u)$ be a multifunction mapping Q to the closed convex subsets of $\mathbb{R}^n$. We assume that*

(a) *For each $t \in [a,b]$, the set*

$$G(t) = \big\{ (u,z) : (t,u,z) \in Q \times \mathbb{R}^n,\ z \in \Gamma(t,u) \big\}$$

is closed and nonempty;

(b) *For every measurable function u on $[a,b]$ satisfying $(t,u(t)) \in Q$ a.e. and every $p \in \mathbb{R}^n$, the support function map*

$$t \mapsto H_{\Gamma(t,u(t))}(p) = \sup \big\{ \langle p, v \rangle : v \in \Gamma\big(t,u(t)\big) \big\}$$

is measurable;

(c) *For a summable function k, we have $\Gamma(t,u) \subset B(0,k(t))\ \forall (t,u) \in Q$.*

Let u_i be a sequence of measurable functions on $[a,b]$ having $(t,u_i(t)) \in Q$ a.e. and converging almost everywhere to u_*, and let $z_i : [a,b] \to \mathbb{R}^n$ be a sequence of functions satisfying $|z_i(t)| \leqslant k(t)$ a.e. whose components converge weakly in $L^1(a,b)$ to those of z_*. Suppose that, for certain measurable subsets Ω_i of $[a,b]$ satisfying $\lim_{i\to\infty} \operatorname{meas} \Omega_i = b - a$, we have

$$z_i(t) \in \Gamma\big(t, u_i(t)\big) + B\big(0, r_i(t)\big),\ \ t \in \Omega_i \text{ a.e.},$$

where r_i is a sequence of nonnegative functions converging in $L^1(a,b)$ to 0. Then we have in the limit $z_*(t) \in \Gamma(t, u_*(t))$, $t \in [a,b]$ a.e.

Proof. Let $H : Q \times \mathbb{R}^n \to \mathbb{R}$ be the support function associated with Γ:

$$H(t,u,p) = \sup \{ \langle p, v \rangle : v \in \Gamma(t,u) \}.$$

Note that $|H(t,u,p)| \leqslant |p| k(t)$, in view of hypothesis (c); it follows that for each $(t,u) \in Q$, the function $p \mapsto H(t,u,p)$ is continuous with respect to p, as the support function of a nonempty bounded set. Furthermore, for any $t \in [a,b]$, using the fact that $G(t)$ is closed, it is not hard to show that for fixed p, the map $u \mapsto H(t,u,p)$ is upper semicontinuous on the set $\{u : (t,u) \in Q\}$ (exercise).

In view of Prop. 2.42, and because Γ is convex-valued, the conclusion that we seek may now be restated as follows: for some null set N, for all $t \in [a,b] \setminus N$, we have

$$H\big(t, u_*(t), p\big) \geqslant \langle p, z_*(t) \rangle \;\; \forall p \in \mathbb{R}^n. \qquad (*)$$

By the continuity of H in p, it is equivalent to obtain this conclusion for all p having rational coordinates. Then, if $(*)$ holds for each such p except on a null set (depending on p), we obtain the required conclusion, since the countable union of null sets is a null set.

We may summarize to this point as follows: it suffices to prove that for any fixed $p \in \mathbb{R}^n$, the inequality in $(*)$ holds almost everywhere.

This assertion in turn would result from knowing that the following inequality holds for any measurable subset A of $[a,b]$:

$$\int_A \big\{ H\big(t, u_*(t), p\big) - \langle p, z_*(t) \rangle \big\}\, dt \;\geqslant 0.$$

(Note that the integrand in this expression is measurable by hypothesis (b), and summable because of the bound on H noted above.) But we have

$$\begin{aligned} \int_A \big\{ H\big(t, u_*(t), p\big) - \langle p, z_*(t) \rangle \big\}\, dt &\geqslant \int_A \Big\{ \limsup_{i \to \infty} H\big(t, u_i(t), p\big) - \langle p, z_*(t) \rangle \Big\}\, dt \\ &\geqslant \limsup_{i \to \infty} \int_A \big\{ H\big(t, u_i(t), p\big) - \langle p, z_i(t) \rangle \big\}\, dt, \end{aligned}$$

as a result of the almost everywhere convergence of u_i to u_*, the upper semicontinuity of H in u, Fatou's lemma, and the weak convergence of z_i to z_*. The last integral above may be written in the form

$$\int_{A \cap \Omega_i} \big\{ H(t, u_i, p) - \langle p, z_i \rangle \big\}\, dt + \int_{A \setminus \Omega_i} \big\{ H(t, u_i, p) - \langle p, z_i \rangle \big\}\, dt. \qquad (**)$$

We have now reduced the proof to showing that the lower limit of this expression is nonnegative.

Using the bound on $|H|$ noted above, together with the given bound on $|z_i|$, we see that the *second* term in (∗∗) is bounded above in absolute value by

$$\int_{[a,b]\setminus\Omega_i} 2|p|k(t)\,dt$$

which tends to 0 as $i \to \infty$, since meas $\Omega_i \to b-a$.

As for the *first* term in (∗∗), the hypotheses imply

$$H\big(t,u_i(t),p\big) \geqslant \langle p,z_i(t)\rangle - r_i(t)|p|, \;\; t \in \Omega_i \text{ a.e.},$$

so that it is bounded below by

$$-\int_{A\cap\Omega_i} r_i(t)|p|\,dt \;\geqslant -\int_a^b r_i(t)|p|\,dt$$

(recall that the functions r_i are nonnegative). But this last term also converges to 0, since r_i converges to 0 in $L^1(a,b)$ by hypothesis. The proof is complete. □

6.40 Exercise. Let A be a compact convex subset of $\mathbb{R}$, and v_i a sequence converging weakly in $L^1(a,b)$ to a limit v_*, where, for each i, we have $v_i(t) \in A$ a.e. Prove that $v_*(t) \in A$ a.e. Show that this may fail when A is not convex. □

In later chapters, Theorem 6.39 will be used when $u : [a,b] \to \mathbb{R}^n$ is absolutely continuous (that is, each component of u belongs to AC$[a,b]$) and $z_i = u_i'$. Furthermore, the convergence hypotheses will most often be obtained with the help of the following well-known result, which we promote to the rank of a theorem:

6.41 Theorem. (Gronwall's lemma) *Let $x : [a,b] \to \mathbb{R}^n$ be absolutely continuous and satisfy*

$$|x'(t)| \leqslant \gamma(t)|x(t)| + \beta(t), \;\; t \in [a,b] \text{ a.e.},$$

where $\gamma, \beta \in L^1(a,b)$, with γ nonnegative. Then, for all $t \in [a,b]$, we have

$$|x(t)-x(a)| \;\leqslant \int_a^t \exp\Big(\int_s^t \gamma(r)\,dr\Big)\big\{\gamma(s)|x(a)| + \beta(s)\big\}\,ds.$$

Proof. Let $r(t) = |x(t)-x(a)|$, a function which is absolutely continuous on $[a,b]$, as the composition of a Lipschitz function and an absolutely continuous one. Let t be in that set of full measure in which both $x'(t)$ and $r'(t)$ exist. If $x(t) \neq x(a)$, we have

$$r'(t) = \Big\langle \frac{x(t)-x(a)}{|x(t)-x(a)|}, x'(t)\Big\rangle,$$

and otherwise $r'(t) = 0$ (since r attains a minimum at t). Thus we have

$$r'(t) \leqslant |x'(t)| \leqslant \gamma(t)|x(t)| + \beta(t) \leqslant \gamma(t)|x(t) - x(a)| + \gamma(t)|x(a)| + \beta(t)$$
$$= \gamma(t)r(t) + \gamma(t)|x(a)| + \beta(t).$$

We may rewrite this inequality in the form

$$\big[r'(t) - \gamma(t)r(t)\big]\exp\Big(-\int_a^t \gamma\Big) \leqslant \exp\Big(-\int_a^t \gamma\Big)\big\{\gamma(t)|x(a)| + \beta(t)\big\}.$$

Note that the left side is the derivative of the function

$$t \mapsto r(t)\exp\Big(-\int_a^t \gamma\Big).$$

With this in mind, integrating both sides of the preceding inequality from a to t yields the required estimate. □

6.42 Exercise. Let $p_i : [a,b] \to \mathbb{R}^n$ be a sequence of absolutely continuous functions with $|p_i(a)|$ uniformly bounded, and such that, for certain functions γ, β in $L^1(a,b)$, we have, for each i:

$$|p_i'(t)| \leqslant \gamma(t)|p_i(t)| + \beta(t), \; t \in [a,b] \text{ a.e.}$$

Then there exist an absolutely continuous function $p : [a,b] \to \mathbb{R}^n$ and a subsequence p_{i_j} such that (componentwise)

$$p_{i_j} \to p \text{ uniformly on } [a,b], \; p_{i_j}' \to p' \text{ weakly in } L^1(a,b).$$ □

Chapter 7
Hilbert spaces

We now pursue our study of Banach spaces in an important special case, one that is characterized by the existence of a certain bilinear function having special properties. Let X be a normed space. A mapping

$$(x,y) \mapsto \langle x,y \rangle_X$$

from $X \times X$ to $\mathbb{R}$ is called *bilinear* if the map $x \mapsto \langle x,y \rangle_X$ is a linear functional for each $y \in X$, as is $y \mapsto \langle x,y \rangle_X$ for each x. We also impose commutativity:

$$\langle x,y \rangle_X = \langle y,x \rangle_X \;\; \forall x,y \in X.$$

A Banach space X is said to be a **Hilbert space** if there is a mapping that has these properties and generates its norm, in the following sense:

$$\|x\|^2 = \langle x,x \rangle_X \quad \forall x \in X.$$

It follows in this case that $\langle x,x \rangle_X \geqslant 0 \;\; \forall x$, with equality if and only if $x = 0$. The bilinear mapping is referred to as an **inner product** on X.

Canonical cases of Hilbert spaces include $\mathbb{R}^n$, $L^2(\Omega)$, and ℓ^2. We have, for example, the following inner products:[1]

$$\langle u,v \rangle_{\mathbb{R}^n} = u \bullet v, \quad \langle f,g \rangle_{L^2(\Omega)} = \int_\Omega f(x)\,g(x)\,dx.$$

It turns out that some rather remarkable consequences for the structure of the space X follow from the mere existence of a scalar product. We suspect that the reader may be familiar with the more immediate ones. Nonetheless, let us begin with a review of the basic theory of Hilbert spaces.

[1] In dealing with vector spaces defined over the complex number field rather than the reals, inner products are correspondingly complex-valued. In that case, the condition $\langle x,y \rangle_X = \overline{\langle y,x \rangle_X}$ is imposed, where the bar refers to complex conjugate. We continue to limit attention throughout, however, to the real-valued case.

F. Clarke, *Functional Analysis, Calculus of Variations and Optimal Control*, Graduate Texts in Mathematics 264, DOI 10.1007/978-1-4471-4820-3_7,

7.1 Basic properties

The first conclusion below is called the **Cauchy-Schwarz inequality**, and the second is known as the **parallelogram identity**.

7.1 Proposition. *Let X be a Hilbert space, and let x, y be points in X. Then*

$$|\langle x,y\rangle_X| \leqslant \|x\|\,\|y\| \quad \textit{and} \quad \left\|\frac{x+y}{2}\right\|^2 + \left\|\frac{x-y}{2}\right\|^2 = \frac{1}{2}\left(\|x\|^2+\|y\|^2\right).$$

Proof. We may suppose $x, y \neq 0$. For any $\lambda > 0$, we have

$$0 \leqslant \|x-\lambda y\|^2 = \langle x-\lambda y, x-\lambda y\rangle_X = \|x\|^2 - 2\lambda\langle x,y\rangle + \lambda^2\|y\|^2.$$

This yields $2\langle x,y\rangle \leqslant \|x\|^2/\lambda + \lambda\|y\|^2$. Putting $\lambda = \|x\|/\|y\|$ gives the required inequality. The identity is proved by writing the norms in terms of the inner product and expanding. □

It follows from the parallelogram identity that the squared norm is strictly convex, in the sense that

$$x \neq y \implies \left\|\frac{x+y}{2}\right\|^2 < \frac{1}{2}\left(\|x\|^2+\|y\|^2\right).$$

In fact, we obtain uniform convexity of the space, as we now see.

7.2 Theorem. *Any Hilbert space is uniformly convex, and therefore reflexive. To every element ζ in X^* there corresponds a unique $u \in X$ such that*

$$\zeta(x) = \langle u, x\rangle_X \;\; \forall x \in X, \;\; \|\zeta\|_* = \|u\|.$$

Proof. Let $\varepsilon > 0$ and $x, y \in B$ satisfy $\|x-y\| > \varepsilon$. It is not difficult to show that by the parallelogram identity we have

$$\left\|\frac{x+y}{2}\right\|^2 < 1-\frac{\varepsilon^2}{4} \implies \left\|\frac{x+y}{2}\right\| < 1-\delta,$$

where $\delta := 1-[1-\varepsilon^2/4]^{1/2}$. This confirms that X is uniformly convex, and therefore reflexive by Theorem 6.3. Let us now consider the linear operator $T : X \to X^*$ defined by

$$\langle Tx, y\rangle = \langle x, y\rangle_X \;\; \forall y \in X.$$

We deduce $\|Tx\|_* = \|x\|$, by the Cauchy-Schwarz inequality. Thus T is norm-preserving, and $T(X)$ is a closed subspace of X^* by Prop. 5.3. To conclude the proof of the theorem, it suffices to prove that $T(X)$ is dense in X^*. Assume the contrary; then there exists $\theta \in X^{**}$ different from 0 such that $\langle \theta, T(X)\rangle = 0$ (see Theorem 2.39). Because X is reflexive, we may write $\theta = J\bar{x}$ for some point $\bar{x} \in X$.

Then

$$0 = \langle \theta, Tx \rangle = \langle J\bar{x}, Tx \rangle = \langle Tx, \bar{x} \rangle = \langle x, \bar{x} \rangle_X \ \ \forall x \in X,$$

whence $\bar{x} = 0$ and $\theta = 0$, a contradiction. □

Convention. The mapping $\zeta \mapsto u$ above is an isometry from X^* to X. It is natural, therefore, to identify the dual X^* of a Hilbert space X with the space itself, which is customary. Having done so, we may drop the symbol X in writing the inner product, since it is equivalent to interpret $\langle x, y \rangle$ as either the inner product, or else the effect of x (viewed as an element of the dual) upon the point y (or vice versa).

Projection. An operation of fundamental importance in Hilbert space is that of projection onto closed convex sets.

7.3 Proposition. *Let C be a closed, convex, nonempty subset of a Hilbert space X, and let $x \in X$. Then there exists a unique $u \in C$ satisfying $d_C(x) = \|x - u\|$.*

Proof. The existence of a closest point u is known (Exer. 5.52). The uniqueness follows from the strict convexity of the squared norm, as follows. If v is a different closest point, then writing the strict convexity inequality for the squared norm (with $u - x$ and $v - x$) yields

$$\left\| \frac{u+v}{2} - x \right\|^2 < \frac{1}{2} \left(\|u - x\|^2 + \|v - x\|^2 \right) = d_C(x)^2,$$

a contradiction, since $(u+v)/2 \in C$. □

The point u is called the **projection** of x onto C, denoted $\operatorname{proj}_C(x)$. We proceed to characterize it geometrically. (Since we identify X^* with X, the normal cone $N_C(u)$ below is naturally viewed as lying in the space X itself.)

7.4 Proposition. *Let C be a closed, convex, nonempty subset of a Hilbert space X, and let $x \in X$, $u \in C$. Then*

$$u = \operatorname{proj}_C(x) \iff x - u \in N_C(u) \iff \langle x - u, y - u \rangle \leqslant 0 \ \ \forall y \in C.$$

Proof. The reader will recall that the last condition in the three-way equivalence is simply a restatement of the fact that $x - u$ is a normal vector in the convex sense (Prop. 2.9). Let us first consider $u = \operatorname{proj}_C(x)$. Let $y \in C$ and $0 < t < 1$. Since C is convex, the fact that u is a closest point allows us to write

$$\|x - u\|^2 \leqslant \|x - (1-t)u - ty\|^2 = \|(x - u) + t(u - y)\|^2.$$

We expand on the right in order to obtain $2t\langle x - u, y - u \rangle \leqslant t^2\|y - u\|^2$. Now divide by t, and then let t decrease to 0; we derive the conclusion $x - u \in N_C(u)$.

Conversely, let $x - u \in N_C(u)$. Then for all $y \in C$, we deduce

$$\|x-u\|^2 - \|y-x\|^2 = 2\langle x-u, y-u\rangle - \|u-y\|^2 \leqslant 0,$$

whence $u = \text{proj}_C(x)$. □

7.5 Exercise. Let C be as in Prop. 7.4. Then

$$\|\text{proj}_C(x) - \text{proj}_C(y)\| \leqslant \|x-y\| \;\; \forall x, y \in X. \qquad \square$$

Another special feature of Hilbert spaces is that subspaces admit complements. For a subset A of X, we define

$$A^\perp = \{\zeta \in X : \langle \zeta, x\rangle = 0 \quad \forall x \in A\}.$$

The closed subspace $A^\perp$ (often pronounced "A perp") is the *orthogonal* to A.

7.6 Exercise. Let M be a subspace of X. Show that $M \cap M^\perp = \{0\}$. □

7.7 Proposition. *Let M be a closed subspace of a Hilbert space X. Then every point $x \in X$ admits a unique representation $x = m + \mu$ where $m \in M$ and $\mu \in M^\perp$; the point m coincides with* $\text{proj}_M(x)$.

Proof. Let $x \in X$, and set $m = \text{proj}_M(x)$. By Prop. 7.4, we have

$$\langle x-m, y-m\rangle \leqslant 0 \;\; \forall y \in M.$$

Since M is a subspace, we deduce from this $\langle x-m, y\rangle = 0 \;\; \forall y \in M$; that is, the point $x-m$ lies in $M^\perp$. Then $x = m + (x-m)$ expresses x in the desired fashion. If $x = m' + \mu'$ is another such decomposition, then

$$m - m' = \mu' - \mu \in M \cap M^\perp = \{0\},$$

whence $m' = m$ and $\mu' = \mu$. The representation is therefore unique. □

7.8 Exercise. In the context of Prop. 7.7, show that the mapping $\text{proj}_M : X \to X$ belongs to $L_C(X, X)$ (thus, projection onto a subspace is a linear operator). □

Orthonormal sets. Prop. 7.7 showed how to decompose a given element x into two components relative to a given closed subspace M, a first component that lies in M, and another that is orthogonal to the first. A more refined use of this decomposition technique can be developed, based on the following concept. Two nonzero points u and v in a Hilbert space X are **orthogonal** if they satisfy $\langle u, v\rangle = 0$. A collection of points $\{u_\alpha : \alpha \in A\}$ in X is said to be *orthonormal* if

$$\|u_\alpha\| = 1 \;\; \forall \alpha, \;\; \langle u_\alpha, u_\beta\rangle = 0 \text{ when } \alpha \neq \beta.$$

7.9 Exercise. Prove that an orthonormal set is independent. □

The following result characterizes projection onto finite-dimensional subspaces.

7.10 Proposition. *Let X be a Hilbert space, and let M be a subspace of X generated by a finite orthonormal set $\{u_i : i = 1,\dots,n\}$. Then*

$$\operatorname{proj}_M(x) = \sum_{i=1}^n \langle x, u_i \rangle u_i, \ \text{ we have } x - \operatorname{proj}_M(x) \in M^\perp,$$

and

$$d_M(x)^2 = \|x\|^2 - \sum_{i=1}^n |\langle x, u_i \rangle|^2.$$

Proof. By expressing the norm in terms of the inner product and expanding, we find

$$\Big\|x - \sum_{i=1}^n \lambda_i u_i\Big\|^2 = \|x\|^2 - 2\sum_{i=1}^n \lambda_i \langle x, u_i \rangle + \sum_{i=1}^n \lambda_i^2.$$

The right side defines a convex function of $\lambda = (\lambda_1, \lambda_2, \dots, \lambda_n)$ which attains a global minimum where its gradient vanishes; that is, for

$$\lambda_i = \langle x, u_i \rangle, \ i = 1,\dots,n.$$

The corresponding linear combination of the u_i is therefore $\operatorname{proj}_M(x)$. We know that $x - \operatorname{proj}_M(x) \in M^\perp$ from Prop. 7.7. Finally, the expression for $d_M(x)^2$ follows from expanding $\|x - \operatorname{proj}_M(x)\|^2$. □

Remark. A finite dimensional subspace M always admits an orthonormal basis $\{u_i : i = 1,\dots,n\}$. For example, if M is of dimension two, let u and v be any two independent elements of M. We set $u_1 = u/\|u\|$, and then we calculate

$$\tilde{v} = v - \langle v, u \rangle u_1.$$

This subtracts from v what is referred to as its *component in the u_1 direction*. Note that $\tilde{v} \neq 0$, since u_1 and v are independent. We proceed to define $u_2 = \tilde{v}/\|\tilde{v}\|$. Then $\{u_1, u_2\}$ is an orthonormal set whose vector span is that of $\{u, v\}$; that is, M. The procedure we have just described can be extended to any finite dimension, and is known as *Gram-Schmidt orthogonalization*.

Hilbert bases. Given an orthonormal set $\{u_\alpha : \alpha \in A\}$ in a Hilbert space X, and a point $x \in X$, we define $\hat{x}_\alpha = \langle x, u_\alpha \rangle$. These are the **Fourier coefficients** of x with respect to the orthonormal set.

7.11 Theorem. (Bessel) *Let $\{u_\alpha : \alpha \in A\}$ be an orthonormal set in a Hilbert space X, and let $x \in X$. Then the set*

$$A(x) = \{\alpha \in A : \hat{x}_\alpha \neq 0\}$$

is finite or countable, and we have

$$\sum_{\alpha \in A} |\hat{x}_\alpha|^2 = \sum_{\alpha \in A(x)} |\hat{x}_\alpha|^2 \leqslant \|x\|^2 \ \forall x \in X.$$

Proof. By definition, the sum on the left is given by

$$\sup\Big\{\sum_{\alpha\in F}|\hat{x}_\alpha|^2 : F\subset A,\ F \text{ finite}\Big\},$$

so the inequality follows from the last assertion of Prop. 7.10 (since $d_M(x)^2 \geqslant 0$). We also deduce that for each i, the set of indices α for which $|\hat{x}_\alpha| > 1/i$ is finite, which implies that $A(x)$ is countable. □

7.12 Exercise. Let $\{u_i : i \geqslant 1\}$ be an orthonormal set in a Hilbert space X. Show that the sequence u_i (considered as a sequence in X^*) converges weakly to 0. Prove that for any $x \in X$, the points

$$S_N = \sum_{i=1}^{N} \hat{x}_i\, u_i$$

define a Cauchy sequence in X. □

A *maximal* orthonormal set $\{u_\alpha : \alpha \in A\}$ is called a **Hilbert basis** for X.

7.13 Exercise. Use Zorn's lemma to prove that a Hilbert basis exists. Show that the distance between any two distinct elements of an orthonormal set in a Hilbert space is $\sqrt{2}$. Deduce that a Hilbert basis for a separable Hilbert space must be finite or countable. □

7.14 Proposition. *If $\{u_\alpha : \alpha \in A\}$ is a Hilbert basis for a Hilbert space X, then the finite linear combinations of the u_α constitute a dense set in X.*

Proof. If the vector space spanned by $\{u_\alpha : \alpha \in A\}$ is not dense in X, then its closure M is such that $M^\perp$ is nontrivial (by Prop. 7.7). This leads directly to a contradiction of the maximality of the orthonormal set $\{u_\alpha : \alpha \in A\}$, as the reader may easily show. □

7.15 Corollary. *A Hilbert space is separable if and only if it admits a Hilbert basis that is finite or countable.*

Proof. If a Hilbert space admits a finite or countable Hilbert basis, then it follows from the proposition that the countable set of finite linear combinations of the basis elements with rational coefficients is dense. The converse follows from Exer. 7.13. □

Remark. Of course, a Hilbert basis is not a basis in the algebraic sense; when X is infinite-dimensional but separable, it admits a countable Hilbert basis, by the corollary above, whereby no countable vector space basis exists (Exer. 8.4). A Hilbert basis allows us to express any point in the space as a (limiting) sum of its orthogonal components, as we now see, but there are in general an infinite number of terms in the sum.

7.16 Proposition. *Let the countable family* $\{u_i : i \geqslant 1\}$ *be a Hilbert basis for a Hilbert space* X*. Then* x *can be recovered from its Fourier coefficients*:

$$x = \sum_{i \geqslant 1} \hat{x}_i\, u_i := \lim_{N \to \infty} \sum_{i=1}^{N} \hat{x}_i\, u_i\,.$$

Proof. The infinite sum $S = \sum_{i \geqslant 1} \hat{x}_i\, u_i$ is well defined as a result of Exer. 7.12. For fixed i, we have

$$\Big\langle \sum_{i=1}^{N} \hat{x}_i\, u_i - x\,, u_i \Big\rangle = 0 \;\; \forall N > i\,,$$

whence $\langle S - x\,, u_i \rangle = 0 \;\; \forall i$; it follows from Prop. 7.14 that $S = x$. □

7.17 Exercise. Prove that the representation for x in Prop. 7.16 is unique:

$$x = \sum_{i \geqslant 1} c_i\, u_i \implies c_i = \hat{x}_i \;\; \forall i \geqslant 1\,.$$ □

A Hilbert space **isomorphism** between two Hilbert spaces X and Y refers to an isometry $T : X \to Y$ that also preserves the inner product:

$$\langle Tx, Tu \rangle_Y = \langle x, u \rangle_X \;\; \forall x, u \in X\,.$$

7.18 Theorem. (Parseval) *Let* X *be a separable Hilbert space, and let* $\{u_i : i \in I\}$ *be a finite or countable Hilbert basis for* X*. Then we have*

$$x = \sum_{i \in I} \hat{x}_i\, u_i\,, \;\; \|x\|^2 = \sum_{i \in I} |\hat{x}_i|^2, \;\; \langle x, y \rangle = \sum_{i \in I} \hat{x}_i\, \hat{y}_i \;\; \forall x, y \in X.$$

X is finite dimensional if and only if X admits a finite Hilbert basis, in which case X is isomorphic as a Hilbert space to $\mathbb{R}^n$ for some integer n. When X is infinite dimensional, then X is isomorphic as a Hilbert space to ℓ^2.

Proof. We merely sketch the lines of the proof (which can be extended to the case of a non countable Hilbert basis). The first assertion is Exer. 7.16; the third one (which implies the second) follows from the identity

$$\Big\langle \sum_{i=1}^{N} \hat{x}_i\, u_i\,, \sum_{i=1}^{N} \hat{y}_i\, u_i \Big\rangle = \sum_{i=1}^{N} \hat{x}_i\, \hat{y}_i\,,$$

where, if the basis is infinite, we pass to the limit as $N \to \infty$ (and use the continuity of the map $(x,y) \mapsto \langle x, y \rangle$).

If X is finite dimensional, then using Gram-Schmidt orthogonalization, we may find a finite orthonormal vector basis for X; it then follows easily that X is isomorphic as a Hilbert space to some $\mathbb{R}^n$. (The converse is evident.) In the infinite dimensional case, there is a countable Hilbert basis $\{u_i : i \geqslant 1\}$ for X. Then the map

$$x \mapsto \big(\langle x, u_1 \rangle, \langle x, u_2 \rangle, \dots\big)$$

defines an isomorphism from X to ℓ^2. □

7.19 Exercise. The goal is to prove the following result, the *Lax-Milgram theorem*. (It is a tool designed to solve certain linear equations in Hilbert space.)

Theorem. *Let $b(u,v)$ be a bilinear form on a Hilbert space X. We suppose that b is continuous, and coercive in the following sense: there exist $c>0$, $C>0$ such that*

$$|b(u,v)| \leqslant C\|u\|\|v\|, \quad b(u,u) \geqslant c\|u\|^2 \quad \forall u,v \in X.$$

Then for any $\varphi \in X$, there exists a unique $u_\varphi \in X$ such that

$$b(u_\varphi, v) = \langle \varphi, v \rangle \quad \forall v \in X.$$

If b is symmetric (that is, if $b(u,v) = b(v,u)$ $\forall u,v \in X$), then u_φ may be characterized as the unique point in X minimizing the function $u \mapsto \frac{1}{2}b(u,u) - \langle \varphi, u \rangle$.

(a) Show that the map $u \mapsto Tu := b(u,\cdot)$ defines an element of $L_C(X,X^*)$ satisfying $\|Tu\|_* \geqslant c\|u\|$.

(b) Prove that TX is closed.

(c) Prove that T is onto, and then deduce the existence and uniqueness of u_φ.

(d) Now let b be symmetric, and define $f(u) = b(u,u)/2$. Show that f is strictly convex, and that $f'(u;v) = b(u,v)$ $\forall u,v$.

(e) Prove that the function $u \mapsto (1/2)b(u,u) - \langle \varphi, u \rangle$ attains a unique minimum over X at a point u_φ. Write Fermat's rule to conclude. □

7.2 A smooth minimization principle

The Banach space of differentiable functions defined below will be used later in the proof of the main result of this section. It combines features of the spaces $C_b(X,\mathbb{R})$ and $\mathrm{Lip}_b(X,X^*)$ that the reader met in §5.1.

7.20 Proposition. *Let X be a normed space. The vector space $C_b^{1,1}(X)$ of bounded continuously differentiable functions $g: X \to \mathbb{R}$ whose derivative g' is bounded and Lipschitz is a Banach space when equipped with the norm*

$$\|g\|_{C_b^{1,1}(X)} = \|g\|_{C_b(X,\mathbb{R})} + \|g'\|_{Lip_b(X,X^*)}.$$

Proof. It is evident that we do have a normed space; it is a matter of verifying that it is complete. Accordingly, let g_n be a Cauchy sequence in $C_b^{1,1}(X)$. Then g_n is Cauchy in the Banach space $C_b(X,\mathbb{R})$, so there exists $g \in C_b(X,\mathbb{R})$ such that

$$\|g_n - g\|_{C_b(X,\mathbb{R})} = \sup_{x \in X} |g_n(x) - g(x)| \to 0.$$

It also follows that the sequence g_n' is Cauchy in the space $\mathrm{Lip}_b(X, X^*)$, which we know to be complete; thus, there also exists $\varphi \in \mathrm{Lip}_b(X, X^*)$ such that

$$\| g_n' - \varphi \|_{\mathrm{Lip}_b(X, X^*)} \to 0.$$

We now claim that g' exists and coincides with φ. Note that the functions g_n' have a common Lipschitz constant L. Let x and u be distinct points in X. Then

$$\begin{aligned} g(u) - g(x) - \langle \varphi(x), u - x \rangle &= \lim_{n \to \infty} g_n(u) - g_n(x) - \langle g_n'(x), u - x \rangle \\ &= \lim_{n \to \infty} \langle g_n'(z_n) - g_n'(x), u - x \rangle, \end{aligned}$$

for some $z_n \in (x, u)$, by the mean value theorem. We deduce

$$| g(u) - g(x) - \langle \varphi(x), u - x \rangle | \leqslant \lim_{n \to \infty} L \| z_n - x \| \| u - x \| \leqslant L \| u - x \|^2.$$

This estimate reveals that $g' = \varphi$, and it follows that $\| g_n - g \|_{C_b^{1,1}(X)} \to 0$. □

Differentiability of the norm. Given a normed space X, we shall refer to the function

$$\theta : X \to \mathbb{R}, \ x \mapsto \theta(x) = \| x \|^2,$$

naturally enough, as the **squared norm** function. It is a crucial property of Hilbert spaces that this function is smooth.

7.21 Proposition. *Let X be a Hilbert space. Then the squared norm function θ is continuously differentiable, with $\theta'(x) = 2x$.*

Proof. We calculate

$$\frac{\theta(x + h) - \theta(x) - \langle 2x, h \rangle}{\| h \|} = \| h \|,$$

which goes to 0 as h does. By definition, then, $\theta'(x) = 2x$. □

It follows from the above (and the chain rule) that, in a Hilbert space, *the norm is differentiable*. (The reader must hear, *sotto voce*, the words "except at the origin" in such a sentence; a norm can never be differentiable at 0, because of positive homogeneity.) In a general Banach space, however, this is not necessarily the case. In fact, a space may admit no equivalent norm which is differentiable.

Bump functions. Recall that the *support* of a function $\varphi : X \to \mathbb{R}$ is the set

$$\operatorname{supp} \varphi = \operatorname{cl} \{ x : \varphi(x) \neq 0 \}.$$

We say that a function φ is a **bump function** when φ is continuous and has nonempty bounded support. Any normed space admits a Lipschitz bump function;

one may take, for example

$$\varphi(x) = \min(0, \|x\| - 1).$$

But does a normed space necessarily admit a *smooth* bump function?

The reader has no doubt seen such bump functions on $\mathbb{R}$ in a calculus course, perhaps the following one:

$$\varphi(t) = \begin{cases} e^{-1/[t^2-1]^2} & \text{if } |t| \leqslant 1 \\ 0 & \text{otherwise.} \end{cases}$$

It is a standard exercise to show that this function is smooth, in fact C^∞. The construction (or others like it) can easily be extended to $\mathbb{R}^n$.

In infinite dimensions, however, the existence of a smooth bump function is problematic. The question turns out to be delicate, and linked very closely to the differentiability properties of the norm (or of some equivalent norm) on the space. We shall prove later, without delving too deeply into the issue, that $L^1(0,1)$ (for example) does not admit a smooth norm or bump function. But Hilbert spaces do not disappoint us in this regard, as we now see.

7.22 Proposition. *If X is a Hilbert space, then $C_b^{1,1}(X)$ contains a bump function.*

Proof. Let $r > 0$, and let $\tau : \mathbb{R} \to \mathbb{R}$ be any twice continuously differentiable function which has compact support in $[-r, r]$ and has $\tau(0) > 0$. Then the function φ defined by

$$\varphi(x) = \tau(\|x\|^2)$$

is a bump function in $C_b^{1,1}(X)$, as is easily seen with the help of Prop. 7.21. Note how the smoothness of the norm on X is essential to this construction. □

The following "smooth minimization principle" is to be compared with Theorem 5.19. One observes that the perturbation term is now provided by a continuously differentiable function, whereas before it exhibited a corner.[2]

7.23 Theorem. *Let X be a Hilbert space, and let $f : X \to \mathbb{R}_\infty$ be proper, lower semicontinuous, and bounded below. Then, for any positive ε, there exists $g \in C_b^{1,1}(X)$ such that*

$$|g(x)| \leqslant \varepsilon, \ \|g'(x)\|_* \leqslant \varepsilon, \ \|g'(x) - g'(y)\|_* \leqslant \varepsilon\|x - y\| \ \forall x, y \in X,$$

and such that $f + g$ attains a minimum over X.

[2] This is a special case of a theorem due to Borwein and Preiss, and also of a general result due to Deville, Godefroy, and Zizler; the proof is taken from [20, p.11].

Proof. Let G be a Banach space of bounded continuous functions $g : X \to \mathbb{R}$ with the following properties:

(a) For every $t > 0$ and $g \in G$, the function $h : x \mapsto h(x) := g(tx)$ belongs to G.

(b) We have $\|g\|_G \geqslant \sup_{x \in X} |g(x)| \;\; \forall g \in G$.

(c) The norm on G is invariant under translation: for every $u \in X$ and $g \in G$, the function $h : x \mapsto h(x) := g(x+u)$ belongs to G and satisfies $\|h\|_G = \|g\|_G$.

(d) G contains a bump function φ.

Note that in view of Props. 7.20 and 7.22, the space $G = C_b^{1,1}(X)$ satisfies these conditions. The proof will show that for any such G, there is a dense set of elements $g \in G$ for which $f+g$ attains a minimum. When G is taken to be $C_b^{1,1}(X)$, we obtain the statement of the theorem.

Consider the set U_n defined by

$$\bigl\{ g \in G : \exists\, x_0 \in X \text{ such that } (f+g)(x_0) < \inf \bigl[(f+g)(x) : x \in X \backslash B(x_0, 1/n) \bigr] \bigr\}.$$

It follows from condition (b) that U_n is an open set in G. We proceed to show that it is also dense. Let $g \in G$ and $\varepsilon > 0$. We wish to exhibit $h \in G$ with $\|h\|_G < \varepsilon$ such that $g+h \in U_n$; that is, such that, for some x_0,

$$(f+g+h)(x_0) < \inf \bigl[(f+g+h)(x) : x \in X \backslash B(x_0, 1/n) \bigr].$$

By properties (c) and (d) above, there is a bump function $\varphi \in G$ such that $\varphi(0) \neq 0$. Replacing $\varphi(x)$ by $\alpha\,\varphi(\tau x)$ for appropriate values of α and $\tau > 0$, we may arrange to have

$$\varphi(0) > 0, \;\; \|\varphi\|_G < \varepsilon, \;\; \varphi(x) = 0 \text{ when } \|x\| \geqslant 1/n.$$

Since $f+g$ is bounded below, we can find $x_0 \in X$ such that

$$(f+g)(x_0) < \inf_X (f+g) + \varphi(0).$$

Let $h(x) = -\varphi(x - x_0)$. By property (a), $h \in G$; we also have $\|h\|_G < \varepsilon$. Moreover,

$$(f+g+h)(x_0) = (f+g)(x_0) - \varphi(0) < \inf_X (f+g).$$

For any $x \in X \backslash B(x_0, 1/n)$, we have

$$(f+g+h)(x) = (f+g)(x) \geqslant \inf_X (f+g) > (f+g+h)(x_0),$$

whence $g+h \in U_n$ as required. This proves that U_n is dense.

Baire's category theorem (see Royden [36]) asserts that, in a complete metric space, the countable intersection of open dense sets is dense. Thus, the set

$$S = \bigcap_{n \geqslant 1} U_n$$

is dense in G. The last step in the proof is to show that for any $g \in S$, the function $f+g$ attains a minimum on X.

Let $x_n \in X$ be such that

$$(f+g)(x_n) < \inf\left[(f+g)(x) : x \in X \backslash B(x_n, 1/n)\right].$$

(Such a point exists because $g \in U_n$.) We claim that $x_p \in B(x_n, 1/n)$ for $p \geqslant n$. If this fails, then, by the choice of x_n, we have

$$(f+g)(x_p) > (f+g)(x_n).$$

However, since $\|x_n - x_p\| > 1/n \geqslant 1/p$, the definition of x_p implies

$$(f+g)(x_n) > (f+g)(x_p),$$

a contradiction which establishes the claim.

It follows from the claim that x_n is a Cauchy sequence, converging to some $\bar{x} \in X$. Invoking the lower semicontinuity of f, we find

$$\begin{aligned}(f+g)(\bar{x}) &\leqslant \liminf_{n\to\infty} (f+g)(x_n) \\ &\leqslant \liminf_{n\to\infty} \left\{ \inf\left[(f+g)(x) : x \in X \backslash B(x_n, 1/n)\right]\right\} \text{ (by the choice of } x_n) \\ &\leqslant \inf\left[(f+g)(x) : x \in X \backslash \{\bar{x}\}\right].\end{aligned}$$

Therefore $f+g$ attains a global minimum at $\bar{x}$. □

We remark that the proof of the theorem applies to any Banach space X whose squared norm function is continuously differentiable, and whose derivative is Lipschitz on bounded sets.

7.24 Exercise. Let X be a Hilbert space, and let $f : X \to \mathbb{R}$ be a twice continuously differentiable function that is bounded below. Prove that for every $\varepsilon > 0$, there is a point z in X that satisfies

$$\|f'(z)\|_* \leqslant \varepsilon \quad \text{and} \quad \langle f''(z)v, v\rangle \geqslant -\varepsilon \|v\|^2 \;\; \forall v \in X. \qquad \square$$

7.3 The proximal subdifferential

The reader has seen the useful notion of subgradient used in the context of convex functions. We now proceed to extend the use of subgradients, in a local fashion, and for functions that are not necessarily convex. The basic concept below and some of its immediate properties can be developed in the setting of any normed space. However, we shall soon see the reasons for limiting attention to Hilbert spaces.

Let X be a normed space, and let $f : X \to \mathbb{R}_\infty$ be given, with $x \in \operatorname{dom} f$.

7.25 Definition. *We say that* $\zeta \in X^*$ *is a* **proximal subgradient** *of* f *at* x *if, for some* $\sigma = \sigma(x,\zeta) \geqslant 0$*, for some neighborhood* $V = V(x,\zeta)$ *of* x*, we have*

$$f(y) - f(x) + \sigma \|y - x\|^2 \geqslant \langle \zeta, y - x \rangle \;\; \forall y \in V.$$

The **proximal subdifferential** *of* f *at* x*, denoted* $\partial_P f(x)$*, is the set of all such* ζ*.*

We are not dealing with convex functions f in this section, but let us note that if f does happen to be convex, then we recover with this definition a familiar construct from convex analysis, namely the subdifferential:[3]

7.26 Proposition. *Let* f *be convex. Then* $\partial_P f(x) = \partial f(x)$.

Proof. It follows directly from the definition of (convex) subgradient that any element ζ of $\partial f(x)$ belongs to $\partial_P f(x)$; the inequality of Def. 7.25 holds with $V = X$ and $\sigma = 0$. We need only show, therefore, that an element $\zeta \in \partial_P f(x)$ belongs to $\partial f(x)$. To do so, note that for such a ζ, the convex function

$$y \mapsto g(y) = f(y) + \sigma \|y - x\|^2 - \langle \zeta, y \rangle$$

attains a local minimum at $y = x$ (by definition of proximal subgradient). By Fermat's rule, we have $0 \in \partial g(x)$. By the sum rule (Theorem 4.10), and because the subdifferential at 0 of the squared norm function is $\{0\}$ (exercise), we obtain $\zeta \in \partial f(x)$, as required. ☐

7.27 Exercise. Let f be a function of two variables as follows: $f(x,y) = g(x - y)$, and let (ζ, ψ) belong to $\partial_P f(x,y)$. Prove that $\zeta + \psi = 0$. ☐

Geometrical interpretation. When f is convex, an element ζ of the subdifferential $\partial f(x)$ satisfies the inequality that appears in Def. 7.25 globally and with $\sigma = 0$. As we know, this corresponds to the epigraph of f having a *supporting hyperplane* at $(x, f(x))$. In the case of a proximal subgradient, however, the proximal subgradient inequality merely asserts that locally, f is bounded below by the function

$$y \mapsto f(x) + \langle \zeta, y - x \rangle - \sigma \|y - x\|^2.$$

The graph of this last function, rather than being a plane, corresponds to a (downward opening) parabola which passes through the point $(x, f(x))$, and which, at the point x, has derivative ζ. Geometrically then, proximal subgradients are the slopes at x of *locally supporting parabolas* to $\operatorname{epi} f$.

---

[3] We remark that, just as one says "f prime" or "dee f" for the derivative of f, the conclusion of Prop. 7.26 is often voiced as follows: dee P f equals dee f (or "curly dee f"); a saving of considerably many syllables.

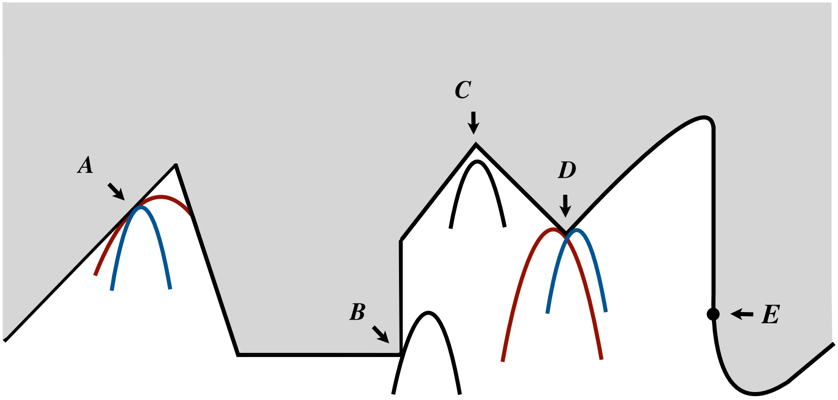


Fig. 7.1
An epigraph, and some locally supporting parabolas.

7.28 Example. Figure 7.1 shows the epigraph of a lower semicontinuous function $f : \mathbb{R} \to \mathbb{R}_\infty$, and focuses on five points in the boundary of epi f.

In a neighborhood of $A = (x_A, f(x_A))$, the function f coincides with the smooth (in fact, affine) function $x \mapsto x + k$, for some constant k. There are infinitely many parabolas that will locally support the epigraph at the point A (two of them are indicated), but they all have slope 1 at x_A; accordingly, we have $\partial_P f(x_A) = \{1\}$.

Near the point $B = (x_B, f(x_B))$, epi f is locally the same as the epigraph of the following function:

$$g(x) = \begin{cases} 0 & \text{if } x \leqslant x_B \\ \infty & \text{if } x > x_B. \end{cases}$$

We find $\partial_P f(x_B) = [0, \infty)$, these being the "contact slopes" of all possible locally supporting parabolas at B.

At the point C, the function is locally of the form $-|x - x_C| + k$, and has a *concave* corner at x_C. Thus, no parabola can locally support epi f at C (curvature will simply not allow it); consequently, we have $\partial_P f(x_C) = \emptyset$.

The point D corresponds to a *convex* corner: f is locally of the form $|x - x_D| + k$. Then $\partial_P f(x_D)$ agrees with the subdifferential $\partial f(x_D) = [-1, 1]$ in the sense of convex analysis. The point E, like C, gives rise to an empty proximal subdifferential, but for a different reason: the infinite slope precludes any supporting parabola. ☐

Relation to derivatives. Let us turn now to the relation between proximal subgradients and derivatives. Suppose that f is Gâteaux differentiable at x. We claim that the only possible element of $\partial_P f(x)$ is $f'_G(x)$. To see this, fix any $v \in \mathbb{R}^n$. Observe

that we may set $y = x+tv$ in the proximal subgradient inequality to obtain

$$(f(x+tv)-f(x))/t \geqslant \langle \zeta, v\rangle - \sigma t\|v\|^2 \text{ for all } t>0 \text{ sufficiently small.}$$

Passing to the limit as $t \downarrow 0$, this yields $\langle f_G'(x), v\rangle \geqslant \langle \zeta, v\rangle$. Since v is arbitrary, we must have $\zeta = f_G'(x)$. We have proved:

7.29 Proposition. *If f is Gâteaux differentiable at x, then $\partial_P f(x) \subset \{f_G'(x)\}$.*

7.30 Example. The last proposition may *fail* to hold with equality; in general, $\partial_P f(x)$ may be empty even when $f_G'(x)$ exists. To develop more insight into this question, the reader should bear in mind that the proximal subdifferential is philosophically linked to (local) convexity, as mentioned above. At points where f has a "concave corner", there will be no proximal subgradients. A simple example is provided by the function $f(x) = -|x|$, where $x \in \mathbb{R}^n$. If $\zeta \in \partial_P f(0)$, then, by definition,

$$-|y| - 0 + \sigma|y|^2 \geqslant \langle \zeta, y\rangle \text{ for all } y \text{ near } 0.$$

Fix any point $v \in X$, and substitute $y = tv$ in the inequality above, for $t > 0$ sufficiently small. Dividing across by t and then letting $t \downarrow 0$ leads to

$$\langle \zeta, v\rangle \leqslant -|v| \;\; \forall\, v \in X,$$

a condition that no ζ can satisfy. Thus, we have $\partial_P f(0) = \emptyset$.

The proximal subdifferential $\partial_P f(x)$ can be empty even when f is continuously differentiable. Consider the function $f(x) = -|x|^{3/2}$ on $\mathbb{R}^n$, which is continuously differentiable with derivative 0 at 0. We claim that $\partial_P f(0) = \emptyset$. To see this, let ζ belong to $\partial_P f(0)$. By Prop. 7.29, ζ must be 0. But then the proximal subgradient inequality becomes

$$-|y|^{3/2} - 0 + \sigma|y|^2 \geqslant 0 \text{ for all } y \text{ near } 0.$$

We let the reader verify that this cannot hold; thus, $\partial_P f(0) = \emptyset$ once again. □

It might be thought that a subdifferential which can be empty is not going to be of much use; for example, its calculus might be very poor. In fact, the possible emptiness of $\partial_P f$ is a *positive* feature in some contexts, as in characterizing certain properties (we shall see this in connection with viscosity solutions later). And the calculus of $\partial_P f$ is complete and rich (but fuzzy, in a way that will be made clear).

It is evident that if f has a (finite) local minimum at x, then $\partial_P f(x)$ is nonempty, since we have $0 \in \partial_P f(x)$ (Fermat's rule). This simple observation will be the key to proving the existence (for certain Banach spaces) of a dense set of points at which $\partial_P f$ is nonempty. First, we require a simple rule in proximal calculus:

7.31 Proposition. *Let $x \in \operatorname{dom} f$, and let $g : X \to \mathbb{R}$ be differentiable in a neighborhood of x, with g' Lipschitz near x. Then*

$$\partial_P(f+g)(x) = \partial_P f(x) + \{g'(x)\}.$$

Proof. We begin with

Lemma. *There exists $\delta > 0$ and M such that*

$$y, z \in B(x,\delta) \implies |g(u) - g(x) - \langle g'(x), u - x \rangle| \leqslant M \|u - x\|^2.$$

To see this, we invoke the Lipschitz hypothesis on g' to find $\delta > 0$ and M such that

$$y, z \in B(x,\delta) \implies \|g'(y) - g'(z)\|_* \leqslant M \|y - z\|.$$

For any $u \in B(x,\delta)$, by the mean value theorem, there exists $z \in B(x,\delta)$ such that $g(u) = g(x) + \langle g'(z), u - x \rangle$. Then, by the Lipschitz condition for g', we have

$$|g(u) - g(x) - \langle g'(x), u - x \rangle| = \\ |\langle g'(z) - g'(x), u - x \rangle| \leqslant M \|z - x\| \|u - x\| \leqslant M \|u - x\|^2,$$

which proves the lemma.

Now let $\zeta \in \partial_P\big(f+g\big)(x)$. Then, for some $\sigma \geqslant 0$ and neighborhood V of x, we have

$$f(y) + g(y) - f(x) - g(x) + \sigma \|y - x\|^2 \geqslant \langle \zeta, y - x \rangle \ \ \forall y \in V.$$

It follows from the lemma that

$$f(y) - f(x) + (\sigma + M)\|y - x\|^2 \geqslant \langle \zeta - g'(x), y - x \rangle \ \ \forall y \in V_\delta := V \cap B(x,\delta),$$

whence $\zeta - g'(x) \in \partial_P f(x)$ by definition. Conversely, if $\psi \in \partial_P f(x)$, then, for some $\sigma \geqslant 0$ and neighborhood V of x, we have

$$f(y) - f(x) + \sigma \|y - x\|^2 \geqslant \langle \psi, y - x \rangle \ \ \forall y \in V.$$

We deduce from the lemma that

$$f(y) + g(y) - f(x) - g(x) + \big(\sigma + M\big)\|y - x\|^2 \geqslant \langle \psi + g'(x), y - x \rangle \ \ \forall y \in V_\delta,$$

whence $\psi + g'(x) \in \partial_P\big(f+g\big)(x)$. □

By taking $f \equiv 0$ in the proposition above, we obtain

7.32 Corollary. *Let $g : X \to \mathbb{R}$ be differentiable in a neighborhood of x, with g' Lipschitz near x. Then $\partial_P g(x) = \{g'(x)\}$.*

7.33 Exercise. Let X and Y be normed spaces. Let $F : X \to Y$ be Lipschitz near x, and $g : Y \to \mathbb{R}$ be C^2 near $F(x)$. Set $f(u) = g(F(u))$. Prove that

$$\partial_P f(x) = \partial_P \langle g'(F(x)), F(\cdot) \rangle (x),$$

where the notation on the right refers to the proximal subdifferential at x of the mapping $u \mapsto \langle g'(F(x)), F(u) \rangle$. (The reader will recognize the chain rule.) □

We now prove that in a Hilbert space, the set of points in $\operatorname{dom} f$ at which at least one proximal subgradient exists is dense in $\operatorname{dom} f$. Minimization will be the key to the proof.

7.34 Theorem. (Proximal density) *Let X be a Hilbert space, and let $f : X \to \mathbb{R}_\infty$ be lower semicontinuous. Let $x \in \operatorname{dom} f$ and $\varepsilon > 0$ be given. Then there exists a point $y \in x + \varepsilon B$ satisfying $\partial_P f(y) \neq \emptyset$ and $|f(y) - f(x)| \leqslant \varepsilon$.*

Proof. By lower semicontinuity, there exists δ with $0 < \delta < \varepsilon$ so that

$$u \in B(x, \delta) \implies f(u) \geqslant f(x) - \varepsilon. \tag{1}$$

We define

$$h(u) = \begin{cases} \left[\, \delta^2 - \|u - x\|^2 \,\right]^{-1} & \text{if } \|u - x\| < \delta, \\ +\infty & \text{otherwise.} \end{cases}$$

Then h is lsc, $h(u) \to \infty$ as u approaches the boundary of $B^\circ(x, \delta)$, h is differentiable on $B^\circ(x, \delta)$, and h' is locally Lipschitz there (see Prop. 7.21). Now consider the function $f + h$, which is lsc and bounded below on $B(x, \delta)$, in fact on X.

Suppose for a moment that X is finite dimensional. Then the function $f + h$ attains a minimum at some point y in $B^\circ(x, \delta)$; thus, $0 \in \partial_P (f + h)(y)$. It now follows from Prop. 7.31 that $-h'(y) \in \partial_P f(y)$, and in particular that $\partial_P f(y) \neq \emptyset$. In view of (1), we may conclude by showing that $f(y) \leqslant f(x)$. We deduce this by noting that y is a minimum of $f + h$, and that $h(x) \leqslant h(y)$, and hence

$$f(y) \leqslant f(x) + \big(h(x) - h(y)\big) \leqslant f(x).$$

The proof is thus complete (and elementary) if X is finite dimensional.

In infinite dimensions, we have to deal with the possible non existence of minimizers. We invoke Theorem 7.23 to deduce that for some $g \in C_b^{1,1}(X)$ having norm less than $\varepsilon/2$, the function $f + h + g$ attains a minimum at a point y. Then

$$0 \in \partial_P (f + h + g)(y) = \partial_P f(y) + h'(y) + g'(y)$$

by Prop. 7.31, whence $\partial_P f(y) \neq \emptyset$. (Note that invoking Theorem 5.19 would not lead to this.) By (1), we have $f(y) \geqslant f(x) - \varepsilon$. In order to conclude, therefore, we need only a corresponding upper bound on $f(y)$. Because $f + h + g$ attains a

minimum at y, we have

$$f(y) \leqslant f(x)+h(x)+g(x)-h(y)-g(y) \leqslant f(x)+g(x)-g(y) < f(x)+\varepsilon. \quad \square$$

The proximal density theorem extends to any Banach space for which (for some equivalent norm) the squared norm function is continuously differentiable, with Lipschitz derivative, on a neighborhood of the origin. (Those are the Hilbert space properties used in the proof.) We close this section by proving that proximal density fails in general, which suggests that proximal calculus is intrinsically limited to Banach spaces having certain smoothness properties.

7.35 Proposition. *There is a Lipschitz function on the Banach space $X = L^1(0,1)$ whose proximal subdifferential is empty at every point.*

Proof. Let $f(x) = -\|x\|_{L^1(0,1)}$. Suppose that $\zeta \in \partial_P f(x)$ for some x, and let us derive a contradiction. We may identify ζ with a function z in $L^\infty(0,1)$, in view of Theorem 6.10. Then, for some σ and $r > 0$, the proximal subgradient inequality of Def. 7.25 yields

$$\int_0^1 |v(t)|\,dt < r \implies \int_0^1 \big\{ |x(t)| - |x(t)+v(t)| - z(t)\,v(t) \big\}\,dt + \sigma \|v\|_1^2 \geqslant 0.$$

In view of the inequality $\|v\|_1^2 \leqslant \int_0^1 |v(t)|^2 dt$, we deduce

$$\int_0^1 |v(t)|\,dt < r \implies \int_0^1 \big\{ |x(t)| - |x(t)+v(t)| - z(t)\,v(t) + \sigma |v(t)|^2 \big\}\,dt \geqslant 0.$$

It now follows from Theorem 6.32 that for almost every $t \in [0,1]$, we have

$$|x(t)| - |x(t)+v| + \sigma |v|^2 - z(t)\,v \geqslant 0 \;\; \forall v \in \mathbb{R}. \tag{2}$$

This cannot hold when $x(t) = 0$ (by the argument given in Example 7.30), so we have $x(t) \neq 0$ a.e. Then the derivative with respect to v of the left side of (2) must vanish, whence $z(t) = -x(t)/|x(t)|$ a.e.

Let S be a set of positive measure such that, for some M, we have $|x(t)| \leqslant M \;\forall t \in S$. For every $\lambda > 0$ sufficiently small, there is a measurable subset S_λ of S having measure λ. We define an element $v \in L^1(0,1)$ by setting $v(t) = -x(t) + z(t)$ when $t \in S_\lambda$, and 0 otherwise. Then $\|v\|_1 \leqslant (M+1)\lambda$, and

$$t \in S_\lambda \implies |x(t)| - |x(t)+v(t)| - z(t)\,v(t) = -2.$$

This, together with the proximal subgradient inequality, implies

$$-2\lambda + \sigma(M+1)^2\lambda^2 \geqslant 0$$

for all $\lambda > 0$ sufficiently small, which provides the desired contradiction. $\square$

In view of the proximal density theorem (as extended to certain Banach spaces), the proposition implies that no equivalent norm on $L^1(0,1)$ exists whose squared norm function has a Lipschitz derivative near 0. The same conclusion holds for $L^\infty(0,1)$ (see Exer. 8.49).

7.36 Exercise. A point $\zeta \in X^*$ is said to be a **proximal supergradient** of f at $x \in \operatorname{dom} f$ if, for some $\sigma = \sigma(x,\zeta) \geqslant 0$, for some neighborhood $V = V(x,\zeta)$ of x, we have

$$f(y) - f(x) - \sigma\|y-x\|^2 \leqslant \langle \zeta, y-x \rangle \;\; \forall y \in V.$$

The set of such ζ, the *proximal superdifferential*, is denoted $\partial^P f(x)$. Show that

$$\partial^P f(x) \,=\, -\partial_P(-f)(x)\,.$$

Suppose that both $\partial_P f(x)$ and $\partial^P f(x)$ are nonempty at x. Prove that f is differentiable at x, and that each of these sets reduces to the singleton $\{f'(x)\}$. □

7.4 Consequences of proximal density

A continuous convex function on a Banach space may fail to be differentiable anywhere (see Exer. 8.49). In Hilbert spaces, however, this cannot happen.

7.37 Proposition. *Let $f : U \to \mathbb{R}$ be convex and lsc on an open convex subset U of a Hilbert space. Then f is differentiable at a dense set of points in U.*

Proof. f is continuous by Theorem 5.17, so that, by Prop. 4.6, $\partial f(x)$ is nonempty for all $x \in U$. Moreover, we have $\partial f(x) = \partial_P f(x)$ by Prop. 7.26. The proximal density theorem, applied to $-f$, implies that $\partial_P(-f)(x) = -\partial^P f(x)$ is nonempty at all points x in a dense subset of U. At these points, f is differentiable by Exer. 7.36. □

The **inf-convolution** of two functions f, g is the function h defined as follows:

$$h(x) \,=\, \inf_{u \in X} \big\{ f(u) + g(x-u) \big\}\,.$$

The term "convolution" is suggested by the visual resemblance of this formula to the classical integral convolution formula. Our interest here involves only such inf-convolutions formed between a function f and the quadratic function $u \mapsto \alpha\|u\|^2$, where $\alpha > 0$. Such functions have surprisingly far-reaching properties, in combination with the proximal density theorem.

Given f, we define $f_\alpha : X \to \mathbb{R}$ by

$$f_\alpha(x) \,=\, \inf_{u \in X} \big\{ f(u) + \alpha\|x-u\|^2 \big\}\,. \tag{1}$$

These functions are known as the *Moreau-Yosida approximations* of f.

7.38 Theorem. *Let X be a Hilbert space, and let $f : X \to \mathbb{R}_\infty$ be proper, lower semicontinuous, and bounded below by a constant c. Then f_α is bounded below by c, and is Lipschitz on each bounded subset of X (and in particular is finite-valued). Furthermore, suppose $x \in X$ is such that $\partial_P f_\alpha(x)$ is nonempty. Then there exists a point $y \in X$ satisfying the following:*

(a) *If u_i is a minimizing sequence for the infimum in (1), then $\lim_{i\to\infty} u_i = y$;*

(b) *The infimum in (1) is attained uniquely at y;*

(c) *The derivative $f'_\alpha(x)$ exists and equals $2\alpha(x-y)$, and $\partial_P f_\alpha(x) = \{2\alpha(x-y)\}$;*

(d) *We have $2\alpha(x-y) \in \partial_P f(y)$.*

Note that in writing conclusions (c) and (d) of the theorem, the dual of the Hilbert space X has been identified with the space X itself.

Proof. Suppose we are given f and $\alpha > 0$ as above. It is clear from the definition that f_α is bounded below by c. We now show that f_α is Lipschitz on any bounded set $S \subset X$. For any fixed $x_0 \in \operatorname{dom} f \neq \emptyset$, note that $f_\alpha(x) \leqslant f(x_0) + \alpha\|x - x_0\|^2$ for all $x \in X$, whence $m := \sup\{f_\alpha(x) : x \in S\} < \infty$. Since $\alpha > 0$, and f is bounded below, we have that, for any $\varepsilon > 0$, the following set is bounded:

$$C = \left\{ z : \exists\, u \in S \text{ such that } f(z) + \alpha\|u - z\|^2 \leqslant m + \varepsilon \right\}.$$

Now let x and u belong to S and $\varepsilon > 0$. Since $f_\alpha(\cdot)$ is given as an infimum, there exists $z \in C$ so that $f_\alpha(u) \geqslant f(z) + \alpha\|u - z\|^2 - \varepsilon$. Thus we have

$$\begin{aligned} f_\alpha(x) - f_\alpha(u) &\leqslant f_\alpha(x) - f(z) - \alpha\|u - z\|^2 + \varepsilon \\ &\leqslant f(z) + \alpha\|x - z\|^2 - f(z) - \alpha\|u - z\|^2 + \varepsilon \\ &= \alpha\|x - u\|^2 - 2\alpha\langle x - u, z - u\rangle + \varepsilon \\ &\leqslant K\|x - u\| + \varepsilon, \text{ where} \end{aligned}$$

$$K = \alpha \sup\left\{ \|s' - s\| + 2\|z - s\| : s', s \in S,\ z \in C \right\} < \infty.$$

Reversing the roles of x and u, and then letting $\varepsilon \downarrow 0$, the above shows that f_α is Lipschitz of rank K on S.

We now consider the other assertions in the theorem. Suppose $x \in X$ is such that there exists at least one $\zeta \in \partial_P f_\alpha(x)$. By the proximal subgradient inequality, there exist positive constants σ and η so that

$$\langle \zeta, u - x\rangle \leqslant f_\alpha(u) - f_\alpha(x) + \sigma\|u - x\|^2 \quad \forall\, u \in B(x, \eta). \tag{2}$$

Now suppose u_i is any minimizing sequence of (1); thus, there exists a sequence $\varepsilon_i \downarrow 0$ such that

$$f_\alpha(x) \leqslant f(u_i)+\alpha\|u_i-x\|^2 = f_\alpha(x)+\varepsilon_i^2. \tag{3}$$

We observe that

$$f_\alpha(u) \leqslant f(u_i)+\alpha\|u_i-u\|^2, \tag{4}$$

since f_α is defined as an infimum over X. Inserting the inequalities (3) and (4) into (2) yields for each $u \in B(x,\eta)$ the conclusion

$$\begin{aligned}\langle \zeta, u-x\rangle &\leqslant \alpha\|u_i-u\|^2-\alpha\|u_i-x\|^2+\varepsilon_i^2+\sigma\|u-x\|^2\\ &= 2\alpha\langle x-u_i, u-x\rangle+\varepsilon_i^2+(\alpha+\sigma)\|u-x\|^2,\end{aligned}$$

which, rewritten, asserts

$$\langle \zeta-2\alpha(x-u_i), u-x\rangle \leqslant \varepsilon_i^2+(\alpha+\sigma)\|u-x\|^2 \;\;\forall u \in B(x,\eta). \tag{5}$$

Now let $v \in B$. Note that $u := x+\varepsilon_i v \in B(x,\eta)$ for large i, since $\varepsilon_i \downarrow 0$. Hence, for all large i, we can invoke (5) to deduce

$$\langle \zeta-2\alpha(x-u_i), v\rangle \leqslant \varepsilon_i(1+\alpha+\sigma).$$

Since $v \in B$ is arbitrary, it follows that $\|\zeta-2\alpha(x-u_i)\| \leqslant \varepsilon_i(1+\alpha+\sigma)$. Set $y = x-\zeta/(2\alpha)$, and observe that (a) immediately follows by letting $i \to \infty$.

To see that y achieves the infimum in (1), it suffices to observe from (a) that

$$f_\alpha(x) \leqslant f(y)+\alpha\|y-x\|^2 \leqslant \liminf_{i\to\infty}\left[f(u_i)+\alpha\|u_i-x\|^2\right] = f_\alpha(x),$$

where the last equality stems from (3). It is also clear that y is unique, since if u is another minimizer of (1), the constant sequence $u_i := u$ is minimizing, and therefore must converge to y by (a). Hence (b) holds.

The following observation about a supergradient (see Exer. 7.36) will be useful in proving the differentiability assertion:

$$\zeta = 2\alpha(x-y) \in \partial^P f_\alpha(x). \tag{6}$$

To prove this, let $u \in X$ and observe that $f_\alpha(u) \leqslant f(y)+\alpha\|u-y\|^2$, with equality holding if $u = x$. Then we see that

$$\begin{aligned}f_\alpha(u)-f_\alpha(x) &\leqslant f(y)+\alpha\|u-y\|^2-f(y)-\alpha\|x-y\|^2\\ &= \langle 2\alpha(x-y), u-x\rangle+\alpha\|u-x\|^2.\end{aligned}$$

This confirms (6), which, in light of Exer. 7.36, implies (c) of the theorem.

As for part (d), observe that the function $u \mapsto f(u)+\alpha\|u-x\|^2$ attains a minimum at $u = y$, so that its proximal subdifferential there contains 0. With the help of Prop. 7.31, this translates to precisely statement (d). □

The distance function. The distance function is destined to play a major role later, when we develop nonsmooth geometry. The following shows that points at which its proximal subdifferential is nonempty have a special character.

7.39 Proposition. *Let S be a nonempty closed subset of a Hilbert space X, and let d_S be its distance function. Let $x \notin S$ be such that $\partial_P d_S(x) \neq \emptyset$. Then* $\operatorname{proj}_S(x)$ *is a singleton $\{s_x\}$, and*

$$\partial_P d_S(x) = \left\{ \frac{x - s_x}{\|x - s_x\|} \right\}.$$

Proof. Consider the function $f = I_S$ and its quadratic inf-convolution

$$f_1(y) = \inf_{u \in X} \left\{ I_S(u) + \|y - u\|^2 \right\} = d_S(y)^2.$$

We have

$$\partial_P f_1(x) = 2\, d_S(x)\, \partial_P d_S(x),$$

according to Exer. 7.33. Thus, $\partial_P f_1(x) \neq \emptyset$, so that by Theorem 7.38, there is a unique point s_x at which the infimum defining $f_1(x)$ is attained. It follows that $\operatorname{proj}_S(x) = \{s_x\}$. Furthermore, Theorem 7.38 asserts that $\partial_P f_1(x)$ is the singleton $2(x - s_x)$. The formula for $\partial_P d_S(x)$ follows from this. □

Closest points. A closed convex set in a Hilbert space admits a unique projection from any point (Prop. 7.3). In general, both the existence and the uniqueness fail in the absence of convexity. However, we have the following density result:

7.40 Exercise. Let S be a nonempty closed subset of a Hilbert space X. Then, for every point x in a dense subset of X, there exists a unique closest point in S to x. □

A linear minimization principle. The following minimization principle is philosophically related to the two others the reader has seen (Theorems 5.19 and 7.23), but it allows the perturbation term to be *linear*.

7.41 Theorem. (Stegall) *Let S be a nonempty, closed, and bounded subset of a Hilbert space X, and let $f : S \to \mathbb{R}_\infty$ be lower semicontinuous and bounded below. Suppose that $S \cap \operatorname{dom} f \neq \emptyset$. Then there exists a dense set of points x in X having the property that the function $u \mapsto f(u) - \langle x, u \rangle$ attains a unique minimum over S.*

Proof. Define

$$g(x) = \inf_{u \in X} \left\{ f(u) + I_S(u) - \tfrac{1}{2}\|u\|^2 + \tfrac{1}{2}\|x - u\|^2 \right\}, \tag{7}$$

which is easily seen to be a function of the form f_α as in (1), where the role of f is played by

$$\tilde{f}(u) := f(u) + I_S(u) - \tfrac{1}{2}\|u\|^2,$$

and where $\alpha = 1/2$. (Note that $\tilde{f}$ is bounded below, in part because S is bounded.) Furthermore, expression (7) for $g(x)$ can be simplified to

$$g(x) = \inf_{u \in S} \{ f(u) - \langle x, u \rangle \} + \tfrac{1}{2} \|x\|^2. \tag{8}$$

It is clear that for fixed $x \in X$, the points u attaining the infima in (7), in (8), as well as in

$$\inf_{u \in S} \{ f(u) - \langle x, u \rangle \} \tag{9}$$

all coincide. The proximal density theorem says that $\partial_P g(x)$ is nonempty for a dense set of points x, and Theorem 7.38 says that for each such x, the infimum in (7) is uniquely attained. Hence, for a dense set of $x \in X$, the infimum in (9) is attained at a unique point in S, which is the assertion of the theorem. □

Chapter 8

Additional exercises for Part I

8.1 Exercise. Give an example of a locally Lipschitz function $f : X \to \mathbb{R}$ defined on a Hilbert space X which is not bounded below on the unit ball. Could such a function be convex? □

8.2 Exercise. Let A be a bounded subset of a normed space X. Prove that

$$\operatorname{co}\big(\partial A\big) \supset \operatorname{cl} A\,. \qquad \square$$

8.3 Exercise. Let X be a normed space, and let A be an open subset having the property that each boundary point x of A admits a supporting hyperplane; that is, there exist $0 \neq \zeta_x \in X^*$ and $c_x \in \mathbb{R}$ such that

$$\langle\, \zeta_x\,, x\,\rangle = c_x\,,\ \ \langle\, \zeta_x\,, u\,\rangle \leqslant c_x \ \ \forall\, u \in A\,.$$

Prove that A is convex. Prove that the result remains valid if the hypothesis "A is open" is replaced by "A is closed and has nonempty interior." □

8.4 Exercise. Let X be an infinite dimensional Banach space. Prove that any vector space basis for X is not countable. By considering ℓ^∞_c, observe that this fact fails for infinite dimensional normed spaces that are not complete. □

8.5 Exercise. Let α_n be a sequence of real numbers, and let $1 \leqslant p \leqslant \infty$. Suppose that, for every $x = (x_1, x_2, \dots)$ in l^p, we have $\sum_{n \geqslant 1} |\alpha_n|\,|x_n| < \infty$. Prove that the sequence α belongs to l^q, where q is the conjugate exponent to p. □

8.6 Exercise. We record a direct definition of the normal cone when S is a subset of $\mathbb{R}^n$, one that does not explicitly invoke polarity to the tangent cone. Let $x \in S$. Show that $\zeta \in N_S(x)$ if and only if, for every $\varepsilon > 0$, there is a neighborhood V of x such that

$$\langle\, \zeta\,, u - x\,\rangle \leqslant \varepsilon\,|\,u - x\,| \ \ \forall\, u \in S \cap V\,. \qquad \square$$

F. Clarke, *Functional Analysis, Calculus of Variations and Optimal Control*, Graduate Texts in Mathematics 264, DOI 10.1007/978-1-4471-4820-3_8,

8.7 Exercise. Let $f : \mathbb{R}^n \to \mathbb{R}$ be given by $f(x) = \langle x, Mx \rangle$, where the matrix M is n by n. Prove that f is convex if and only if $f(x) \geqslant 0 \;\; \forall x$. □

8.8 Exercise. Let X be a normed space, and let C, D be closed convex subsets of X. Show by a counterexample in $\mathbb{R}^2$ that $C+D$ may not be closed. Prove that $C+D$ is closed if one of C or D is compact. □

8.9 Exercise. Let X be a normed space.

(a) Let Σ and Δ be bounded, convex, weak*closed subsets of X^*. Prove that $\Sigma + \Delta$ has the same properties.

(b) Let Σ_i be bounded, convex, weak*closed subsets of X^*, $i = 1, 2, \ldots, n$. Prove that the set $\operatorname{co}\left\{ \bigcup_{i=1}^{n} \Sigma_i \right\}$ is weak*closed. □

8.10 Exercise. Let $f \colon X \times Y \to \mathbb{R}_\infty$ be a convex function, where X, Y are vector spaces. If, for every $x \in X$, we have $g(x) := \inf_{y \in Y} f(x,y) > -\infty$, then prove that g is convex. □

8.11 Exercise. Let $\zeta : X \to \mathbb{R}$ be a nonzero linear functional on a normed space X. Prove that the following are equivalent:

(a) ζ is continuous;

(b) The null space $N(\zeta) := \left\{ x \in X : \langle \zeta, x \rangle = 0 \right\}$ is closed;

(c) $N(\zeta)$ is not dense in X.

Deduce from this that if ζ is a discontinuous linear functional on X, then its null space is dense. □

8.12 Exercise. Construct a Lipschitz function $f \colon \mathbb{R} \to \mathbb{R}$ such that $f'(0;1)$ fails to exist. □

8.13 Exercise. (von Neumann) Let $X = \ell^p$, $1 < p < \infty$, and denote by e_n (as usual) the element of X whose n-th term is 1 and whose other terms are all 0. Then, as we know, the sequence e_n converges weakly, but not strongly, to 0 (Example 3.5).

(a) Prove that the set $A := \left\{ e_n + n\, e_m : m > n \geqslant 1 \right\}$ is strongly closed in X.

(b) Show that any weak neighborhood of 0 contains infinitely many elements of A, but no sequence in A converges weakly to 0.

Deduce that the set of all weak limits of sequences in A fails to be weakly closed; thus, the weak closure of A is not obtained by taking all limits of weakly convergent sequences. (This phenomenon cannot occur when a topology is metrizable.) □

8.14 Exercise. Let x_n be a sequence in ℓ^p $(1 < p < \infty)$, where we write

$$x_n = (x_{n,1}, x_{n,2}, x_{n,3} \dots).$$

Prove that x_n converges weakly to 0 in ℓ^p if and only if the sequence x_n is bounded in ℓ^p and, for each i, we have $\lim_{n\to\infty} x_{n,i} = 0$. □

8.15 Exercise. Prove that in ℓ^1, a sequence converges weakly if and only if it converges strongly. □

8.16 Exercise. Let $X = \ell^\infty$, and let C consist of those points $(x_1, x_2, \dots)$ in X for which $x_i \in [0,1]$ $\forall i$ and $\lim_{i\to\infty} x_i = 1$. Prove that C is a convex, weakly closed, bounded subset of X, but that C is not weak* closed. □

8.17 Exercise. Let $f\colon X \to \mathbb{R}_\infty$ be convex and lsc, where X is a normed space, and suppose that $\lim_{\|u\|\to\infty} f(u) = \infty$. Prove the existence of $\alpha > 0$ and β such that

$$f(u) \geqslant \alpha\|u\| - \beta \quad \forall u \in X. \qquad \square$$

8.18 Exercise. Let X be a normed space.

(a) Let C be a closed subset of X such that $x, y \in C \implies (x+y)/2 \in C$. Prove that C is convex.

(b) Let $f : X \to \mathbb{R}_\infty$ be a lower semicontinuous function such that

$$f\big((x+y)/2\big) \leqslant \tfrac{1}{2} f(x) + \tfrac{1}{2} f(y) \quad \forall x, y \in X.$$

Prove that f is convex. □

8.19 Exercise. Let $f\colon \mathbb{R}^n \to \mathbb{R}$ be convex and differentiable, and suppose that, for certain positive constants a and b, we have

$$0 \leqslant f(x) \leqslant a + b|x|^2 \quad \forall x \in \mathbb{R}^n.$$

Identify constants c and d such that $|\nabla f(x)| \leqslant c + d|x|$ $\forall x \in \mathbb{R}^n$. □

8.20 Exercise. Let $g_i\colon \mathbb{R}^n \to \mathbb{R}$ be convex and differentiable $(i = 1, 2, \dots, m)$, and set $f(x) = \max\{g_i(x) : 1 \leqslant i \leqslant m\}$. We define

$$I(x) = \big\{\text{the indices } i \in \{1, 2, \dots, m\} \text{ such that } g_i(x) = f(x)\big\}.$$

Prove that $\partial f(x) = \operatorname{co}\{g_i'(x) : i \in I(x)\}$. □

8.21 Exercise. Construct a convex function $f : X \to \mathbb{R}_\infty$ on a Hilbert space X such that $\partial f(0)$ is a singleton yet f fails to be Gâteaux differentiable at 0. □

8.22 Exercise. We are interested in minimizing over $\mathbb{R}^3$ the function

$$f(x) \;=\; \tfrac{1}{2}\langle x, Qx\rangle + \langle b, x\rangle \text{ subject to } -1 \leqslant x_i \leqslant 1,\; i = 1,2,3,$$

where

$$Q = \begin{bmatrix} 13 & 12 & -2 \\ 12 & 17 & 6 \\ -2 & 6 & 12 \end{bmatrix}, \quad b = \begin{bmatrix} -22.0 \\ -14.5 \\ 13.0 \end{bmatrix}.$$

Prove that the minimum is attained at the point $x_* = (1, 1/2, -1)$. (Given: the characteristic polynomial of Q is $-\lambda^3 + 42\lambda^2 - 397\lambda + 100$.) □

8.23 Exercise. Let C and D be closed convex subsets of a normed space X such that $\operatorname{int}(C-D) \neq \emptyset$. The goal is to prove that, for any point x in $C \cap D$, we have

$$N_{C\cap D}(x) \;=\; N_C(x) + N_D(x).$$

(Note how this sharpens the result of Exer. 4.11.) We may reduce to the case $x = 0$. Prove the inclusion $\supset$. Now, let $\zeta \in N_{C\cap D}(0)$. Show that we can separate $(0,0)$ from the set

$$\operatorname{int}\big\{\big(c-d,\, \delta - \langle \zeta, d\rangle\big) : \delta \geqslant 0,\, c \in C,\, d \in D\big\},$$

and conclude. □

8.24 Exercise. (Gâteaux differentiability) Let $f : X \to \mathbb{R}$, where X is a normed space.

(a) Give an example with $X = \mathbb{R}^2$ in which f is discontinuous at the origin yet Gâteaux differentiable there.

(b) Suppose that f is Gâteaux differentiable at each point u in a neighborhood of x. Suppose in addition that the map $u \to f'_G(u)$ is continuous at x. Prove that f is differentiable at x.

(c) Let $X = \mathbb{R}^n$, and let f be Gâteaux differentiable at x and Lipschitz on a neighborhood of x. Prove that f is differentiable at x. □

8.25 Exercise. Let X be a normed space whose dual ball is strictly convex:

$$\zeta_1, \zeta_2 \in B_*,\; \zeta_1 \neq \zeta_2 \implies \big\|(\zeta_1 + \zeta_2)/2\big\|_* < 1.$$

Prove that $x \mapsto \|x\|$ is Gâteaux differentiable at every nonzero point. □

8.26 Exercise. Let U be an open convex subset of $\mathbb{R}^n$, and let $f : U \to \mathbb{R}$ be C^2. Suppose that for all x in U with the exception of at most countably many, the Hessian matrix $\nabla^2 f(x)$ has strictly positive eigenvalues. Prove that f is strictly convex. □

8.27 Exercise. Let $f : X \to \mathbb{R}$ be convex, where X is a normed space. Show that the multifunction $\partial f(\cdot)$ is *monotone*:

$$\langle \zeta_1 - \zeta_2, x_1 - x_2 \rangle \geqslant 0 \;\; \forall x_1, x_2 \in X, \; \zeta_i \in \partial f(x_i) \; (i = 1, 2).$$

Prove that if $f : X \to \mathbb{R}$ is a Gâteaux differentiable function satisfying

$$\langle f'_G(x_1) - f'_G(x_2), x_1 - x_2 \rangle \geqslant 0 \;\; \forall x_1, x_2 \in X,$$

then f is convex. □

8.28 Exercise. Let $T : X \to Y$ be a continuous linear operator, where X and Y are Banach spaces. The goal is to prove the following characterization of surjectivity. (T^* refers to the adjoint of T.)

Theorem. *The following are equivalent*:

1) *T is surjective*;
2) *For some $\delta > 0$, we have $TB_X \supset \delta B_Y$*;
3) *For some $\delta > 0$, we have $\| T^* y_* \|_{X^*} \geqslant \delta \| y_* \|_{Y^*} \;\; \forall y_* \in Y^*$.*

The proof is to be carried out by means of the following steps.

(a) Prove that (1) and (2) are equivalent.

(b) Prove that (2) implies (3).

(c) We assume now that (3) holds, and we prove that T is surjective. We reason by the absurd. Let $y \in Y$ be a point which is not in the range $T(X)$ of T. Show that, for any $\varepsilon > 0$, there exists $x_\varepsilon \in X$ which minimizes over X the function $h(x) = \| Tx - y \|_Y + \varepsilon \| x - x_\varepsilon \|_X$.

(d) Apply subdifferential calculus to deduce that $Tx_\varepsilon = y$, provided ε has been chosen sufficiently small. This contradiction completes the proof. □

8.29 Exercise. (Steiner points) Let k distinct points $x_1, x_2, \ldots, x_k$ in $\mathbb{R}^n$ be given ($k \geqslant 3$), and consider the problem of finding a point that is "central" with respect to them, by which we mean that we seek $x \in \mathbb{R}^n$ that minimizes the function

$$f(x) = \sum_{i=1}^{k} |x - x_i|$$

(the sum of the distances) over $\mathbb{R}^n$. We refer to a point x_* that minimizes f as a *Steiner point*.

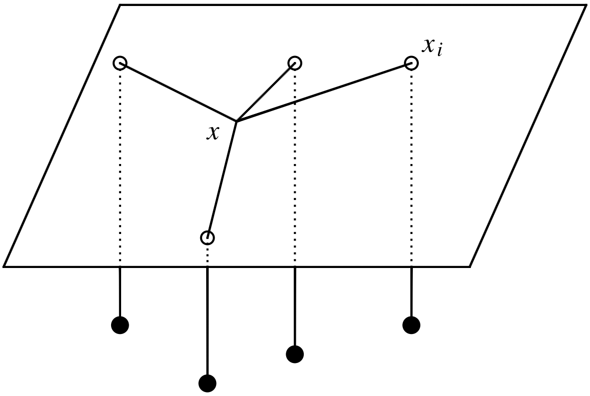
(a) Prove that the set of Steiner points is nonempty, convex, and compact.

(b) A natural question is whether one of the points x_j can itself be a Steiner point. For given $j \in \{1, 2, \ldots, k\}$, prove that x_j is a Steiner point if and only if

$$\Big| \sum_{i \neq j} \frac{x_j - x_i}{|x_j - x_i|} \Big| \leqslant 1 .$$

(c) The case $n = 2, k = 3$ was considered by Torricelli (a contemporary of Galileo). Let $\theta \in [0, \pi]$ be the angle formed by the segments $[x_1, x_2]$ and $[x_2, x_3]$. Prove his theorem stating that (in the current terminology) x_2 is a Steiner point if and only if $\theta \geqslant 2\pi/3$.

(d) When $n = 2$, a Steiner point can be found mechanically, as follows. Pass strings through k holes in a table, the holes being located at the points x_i; let unit masses hang from each string under the table, and join all the strings together on the table top at a nexus point x (see the figure below). Explain why (in the absence of friction, and using massless strings) the nexus will then move to a Steiner point.



(e) Find all the Steiner points when $n = 2$ and the $k = 4$ points involved are

$$(1,1),\ (-1,1),\ (-1,-1),\ (1,-1).$$ □

8.30 Exercise. Let $f : X \to \mathbb{R}_\infty$ be convex and lower semicontinuous, where X is a normed space. Prove that f is bounded below on bounded sets. □

8.31 Exercise. It turns out that differentiability and strict convexity are dual properties relative to conjugacy, as we now see.

(a) Let $f : \mathbb{R}^n \to \mathbb{R}$ be strictly convex, as well as coercive: $\lim_{|x| \to \infty} f(x)/|x| = \infty$. Prove that f^* is continuously differentiable.

(b) Let $g : \mathbb{R}^n \to \mathbb{R}$ be convex, coercive, and differentiable. Prove that g^* is strictly convex. □

8.32 Exercise. Let $f \colon E \to [0, \infty]$ be proper and lsc, where (E, d) is a complete metric space. Let $g : E \to E$ be a function such that

$$f\big(g(x)\big)+d\big(x,g(x)\big) \leqslant f(x) \quad \forall x \in E.$$

Prove that g admits a fixed point. (This is *Caristi's fixed point theorem.*) □

8.33 Exercise. Prove that the set of operators $T \in L_C(X,X)$ for which T^{-1} exists and belongs to $L_C(X,X)$ is open, and that the mapping $T \mapsto T^{-1}$ is continuous. [Hint. If T is small, then a well-known series shows that $(I-T)^{-1}$ exists.] □

8.34 Exercise. Let X and Y be Banach spaces, and let $T : X \to Y$ be a continuous linear operator which is surjective. Let $g : X \to Y$ be continuously differentiable. We study the solutions of the equation

$$Tx + rg(x) = 0, \tag{$*$}$$

where r is a real parameter. Prove the existence of two positive numbers δ and K having the property that, for every $r \in [-\delta,\delta]$, there exists a solution x_r of ($*$) satisfying

$$\|x_r\|_X \leqslant K|r|,$$

and such that (letting $N(T)$ be the null space of T) we have

$$d\big(x,N(T)\big) \leqslant K\|Tx\|_Y \quad \forall x \in X.$$

□

8.35 Exercise. Let $f : \mathbb{R}^n \to \mathbb{R}$ be convex. Prove the existence of a measurable function $\zeta : \mathbb{R}^n \to \mathbb{R}^n$ such that $\zeta(x) \in \partial f(x)$ a.e. □

8.36 Exercise. Let $f : \mathbb{R}^n \to \mathbb{R}$ be continuous and have superlinear growth:

$$\lim_{|u|\to\infty} f(u)/|u| = \infty.$$

For $x \in \mathbb{R}^n$, let the set of points in $\mathbb{R}^n$ which minimize the function $u \mapsto f(u) - u \bullet x$ be denoted by $\Gamma(x)$. Prove that Γ admits a measurable selection. □

8.37 Exercise. (Inverse function theorem) The goal is to prove Theorem 5.38.

(a) Prove the existence of $\eta > 0$ such that $F'(\bar{x})^* B_* \supset 2\eta B_*$. [Hint. Prove that $F'(\bar{x})^*$ is continuous and surjective.]

(b) Justify the existence of a neighborhood A of $\bar{x}$ such that

$$x \in A \implies \big[F'(x) - F'(\bar{x})\big]B \subset \eta B.$$

(c) Prove that

$$x,u \in A \implies \|F(x) - F(u)\| \geq \eta\|x-u\|.$$

[Hint. Let $\zeta \in B_*$ satisfy $\langle \zeta, x-u \rangle = \|x-u\|$, and then let $\psi \in B_*$ satisfy $F'(\bar{x})^*\psi = 2\eta\zeta$ (ψ exists by (a)). Then apply the mean value theorem to the function $z \mapsto \langle \psi, F(z) \rangle$ on the interval $[x,u]$.]

(d) Invoke Theorem 5.32 to deduce the existence of a neighborhood W of $\bar{y}$ such that, for each $y \in W$, the equation $F(x) = y$ has a unique solution $\hat{x}(y) \in A$.

(e) Prove that $\hat{x}$ is Lipschitz (one may need to reduce W).

(f) Prove that $D\hat{x}(\bar{y})$ exists and equals $F'(\bar{x})^{-1}$. [Hint. Write

$$F(\hat{x}(y)) = F(\bar{x}) + F'(\bar{x})\big(\hat{x}(y) - \bar{x}\big) + o\big(\hat{x}(y) - \bar{x}\big)$$

and show that the last term is of the form $o(y - \bar{y})$.]

(g) Prove that $\hat{x}$ is continuously differentiable near $\bar{y}$, and complete the proof. [Hint. Obtain the formula of the preceding step for y near $\bar{y}$, and use Exer. 8.33.] □

8.38 Exercise. Let S be a nonempty closed subset of a Hilbert space X, with $S \neq X$. Prove the existence of a point $x \in X$ at which the distance function d_S fails to be differentiable. □

8.39 Exercise. (Motzkin's theorem) Let S be a nonempty closed subset of $\mathbb{R}^n$, and for any $x \in S$, let $\text{proj}_S(x)$ denote as usual the (nonempty) set of points $u \in S$ satisfying $d_S(x) = |u - x|$. We prove the following theorem due to Motzkin:

S is convex if and only if $\text{proj}_S(x)$ is a singleton for each x.

The necessity is known to us (see Prop. 7.3), so we turn to the hard part: showing that the uniqueness of closest points implies the convexity of S. We denote by s_x the unique projection of x onto S.

(a) For fixed x, v, let p_t be the projection of $x + tv$ on S. Prove that $\lim_{t \downarrow 0} p_t = s_x$.

(b) Show that for $t > 0$ we have

$$|x + tv - p_t|^2 - |x - p_t|^2 \leqslant d_S^2(x + tv) - d_S^2(x) \leqslant |x + tv - s_x|^2 - |x - s_x|^2.$$

Deduce that the function $d_S^2(x)$ is Gâteaux differentiable at x, with derivative $2(x - s_x)$.

(c) Prove that the function $\varphi(x) = \big(|x|^2 - d_S^2(x)\big)/2$ is convex.

(d) Let $f(x) = |x|^2/2 + I_S(x)$. Prove that $f^* = \varphi$.

(e) We set $g = f^{**} = \varphi^*$, and we prove that $\text{dom}\, g \supset \text{co}\, S$. To this end, note that

$$\begin{aligned} g(x) &= \varphi^*(x) = \sup_{y \in \mathbb{R}^n} \big\{ x \bullet y - |y|^2/2 + d_S(y)^2/2 \big\} \\ &= \sup_{y \in \mathbb{R}^n} \inf_{s \in S} \big\{ x \bullet y - |y|^2/2 + |y - s|^2/2 \big\} \\ &\leqslant \inf_{s \in S} \sup_{y \in \mathbb{R}^n} \big\{ x \bullet y - |y|^2/2 + |y - s|^2/2 \big\} \\ &= \inf_{s \in S} \sup_{y \in \mathbb{R}^n} \big\{ (x - s) \bullet y + |s|^2/2 \big\}, \end{aligned}$$

which equals $+\infty$ if $x \notin S$, and $|x|^2/2$ otherwise. We deduce $\operatorname{dom} g \supset S$, whence $\operatorname{dom} g \supset \operatorname{co} S$.

(f) Let x be a point for which $\partial g(x) \neq \emptyset$. Show that $x \in S$. [Hint: subdifferential inversion.]

(g) Let A be the set of points x such that $\partial g(x) \neq \emptyset$. Show that

$$S \supset \overline{A} \supset \operatorname{dom} g \supset \operatorname{co} S \supset S.$$

This implies $S = \operatorname{co} S$, which reveals that S is convex. □

8.40 Exercise. Let (E,d) be a complete metric space. Given two points u,v in E, the *open interval* (u,v) refers to the set (possibly empty) of points x different from u or v such that $d(u,v) = d(u,x) + d(x,v)$. Let $g : E \to E$ be a continuous mapping satisfying, for a certain $c \in [0,1)$:

$$v \in E,\, v \neq g(v) \implies \exists\, w \in \big(v, g(v)\big) \text{ such that } d\big(g(v), g(w)\big) \leqslant c\, d(v,w).$$

Then g admits a fixed point. □

8.41 Exercise. Let K be a compact subset of $\mathbb{R}^n$ containing at least two points. Show that $C(K)$ is not uniformly convex. □

8.42 Exercise. Prove that the dual space c_0^* of c_0 is isometric to ℓ^1. It follows that c_0 is not reflexive. Find an element in the set $c_0^{**} \setminus (J c_0)$. □

8.43 Exercise. Let X be a normed space. We show that if the weak topology of X admits a countable base of open sets at 0, then X is finite dimensional. (Thus, the weak topology on an infinite dimensional normed space is never metrizable.) Suppose that such a countable base does exist.

(a) Prove the existence of a countable set $\{\zeta_n\}$ in X^* such that every $\zeta \in X^*$ is a finite linear combination of the ζ_n.

(b) Derive from this that X^* is finite dimensional, and then that X is finite dimensional. □

8.44 Exercise. Let X be a uniformly convex Banach space, and let x_n be a sequence converging weakly to x. Prove that if $\limsup_{n\to\infty} \|x_n\| \leqslant \|x\|$, then x_n converges strongly to x. □

8.45 Exercise. Let S be a nonempty subset of $\mathbb{R}^n$, and set $X = L^2(0,1)^n$. We define a subset A of X as follows:

$$A = \big\{ f \in X : f(t) \in S,\ t \in [0,1] \text{ a.e.} \big\}.$$

Prove that A is closed if and only if S is closed, that A is convex if and only if S is convex, and that A is weakly compact if and only if S is compact and convex. □

8.46 Exercise. Let $g : \mathbb{R} \to \mathbb{R}$ be continuous and bounded below, and define

$$f(v) = \int_a^b g(v(t))\,dt\,, \quad v \in L^1(0,1)\,.$$

Show that f is lower semicontinuous. Show that f is weakly lower semicontinuous if and only if g is convex. □

8.47 Exercise. Prove that if a Banach space X is reflexive, then a subset of X^* is weak* closed if and only if it is weakly closed. □

8.48 Exercise. Let X be a Banach space.

(a) Let σ and ζ be nonzero points in X^* such that ζ does not lie on the ray $\mathbb{R}_+\sigma$. Prove the existence of $x \in X$ such that $\langle \zeta, x \rangle < 0 < \langle \sigma, x \rangle$.

(b) Deduce that the ray is weak* closed.

(c) Prove that a convex cone Σ in X^* is weak* closed if and only if the set $\Sigma \cap B_*$ is weak* closed. □

8.49 Exercise. Show that the norm on the Banach space $X = L^\infty(0,1)$ fails to be Gâteaux differentiable at any point. Deduce from this that the proximal subdifferential $\partial_P f$ of the function $f(x) = -\|x\|$ is empty at every point. □

8.50 Exercise. Characterize the (dense) set of points at which the (usual supremum) norm on $C[0,1]$ is Gâteaux differentiable. Are there any points at which the norm is (Fréchet) differentiable? □

8.51 Exercise. Let X be a reflexive Banach space, and let S and Σ be closed, convex, nonempty subsets of X and X^* respectively, at least one of which is bounded. Then

$$\inf_{x \in S} \sup_{\sigma \in \Sigma} \langle \sigma, x \rangle = \sup_{\sigma \in \Sigma} \inf_{x \in S} \langle \sigma, x \rangle\,.$$

□

8.52 Exercise. Let X be a Banach space, and $q : X \times X \to \mathbb{R}$ a function such that, for each $(u,v) \in X \times X$, the mappings $w \mapsto q(u,w)$ and $w \mapsto q(w,v)$ belong to X^*. Thus, $q(u,v)$ is bilinear and continuous with respect to each of its variables.

(a) Prove the existence of M such that $|q(u,v)| \leqslant M\|u\|\|v\| \ \forall (u,v) \in X \times X$.

(b) Prove that the function $f(x) = q(x,x)$ is Gâteaux differentiable at every x, and find an expression for the directional derivative $f'(x;v)$ in terms of q.

We now suppose that the bilinear form q is *symmetric* and *coercive*:

$$q(u,v) = q(v,u) \ \forall (u,v)\,, \quad \exists\, c > 0 \text{ such that } f(x) = q(x,x) \geqslant c\|x\|^2 \ \forall x \in X.$$

(c) Prove that f is strictly convex and continuous.

We now suppose that X is reflexive. Let K be a nonempty closed convex subset of X, and $\zeta \in X^*$.

(d) Prove the existence of a unique point x_* which minimizes over K the function $x \mapsto \frac{1}{2} f(x) - \langle \zeta, x \rangle$.

(e) Deduce that x_* satisfies the necessary condition $-\frac{1}{2} f_G'(x_*) + \zeta \in N_K(x_*)$, and that this is equivalent to the following *variational inequality*:

$$q(x_*, u - x_*) \geqslant \langle \zeta, u - x_* \rangle \ \forall u \in K.$$

(f) Prove that x_* is the unique solution of the variational inequality (which arises notably in certain problems in elasticity).

(The last three assertions together constitute *Stampacchia's theorem on variational inequalities*.) □

8.53 Exercise. Show by an example that Theorem 7.23 fails to hold in an arbitrary Banach space. □

8.54 Exercise. The goal here is to prove a special case of a theorem due to Borwein and Preiss, a minimization principle in which the perturbation term is differentiable, as it is in Theorem 7.23, but having additional features. The result involves two parameters that serve to relate the conclusion to a *pre-given* point x of interest; in that sense, it resembles Theorem 5.19.

Theorem. *Let $f : X \to \mathbb{R}_\infty$ be lsc and bounded below, where X is a Hilbert space. Let $\varepsilon > 0$. Suppose that x is a point satisfying $f(x) < \inf_X f + \varepsilon$. Then, for any $\lambda > 0$ there exist points y and z with*

$$\|z - x\| \leqslant \lambda, \quad \|y - z\| \leqslant \lambda, \quad f(y) \leqslant f(x) + \lambda,$$

and having the property that the function

$$w \mapsto f(w) + \frac{2\varepsilon}{\lambda^2} \|w - z\|^2$$

has a unique minimum at $w = y$.

(a) Consider the inf-convolution f_α (as in Theorem 7.38) with $\alpha = 2\varepsilon/\lambda^2$:

$$f_\alpha(u) = \inf_{w \in X} \left\{ f(w) + \frac{2\varepsilon}{\lambda^2} \|w - u\|^2 \right\}.$$

Prove the existence of $z \in x + \lambda B$ satisfying

$$f_\alpha(z) \leqslant f_\alpha(x) + \min(\varepsilon, \lambda), \ \partial_P f_\alpha(z) \neq \emptyset.$$

(b) Prove that there is a unique point y at which the infimum defining $f_\alpha(z)$ is attained, and that $f(y) \leqslant f(x)+\lambda$.

(c) Prove that $\|y-z\| \leqslant \lambda$. □

Extreme points. Let K be a nonempty subset of a normed space X. An *extreme point* of K refers to a point $x \in K$ which cannot be written in the form $(1-t)\,y+t\,z$, where $t \in (0,1)$ and y, z are distinct points in K. (In other words, x fails to lie in any open segment determined by two points of K.)

8.55 Exercise. Let K be convex, and let $D \subset K$ be such that $\operatorname{co} D = K$. Prove that D contains all the extreme points of K. □

8.56 Exercise. The preceding exercise implies that in seeking subsets of K that generate K by convexification, we must include the extreme points. Will these suffice? The following result provides a positive answer, when K is compact.

Theorem. (Krein-Milman) *If K is a compact convex subset of X, and if E is the set of its extreme points, then we have $K = \overline{\operatorname{co}}\, E$.*

Proof. [1] We extend the concept of extreme point to subsets S of K. We say that S is an *extreme set* if S is nonempty and

$$x \in K,\ y \in K,\ 0 < t < 1,\ (1-t)x+ty \in S \implies x \in S,\ y \in S.$$

Let P denote the family of all compact extreme sets; then P is nonempty, since $K \in P$. Prove the following two lemmas:

Lemma 1. *Let $\{S_\alpha\}$ be a collection of extreme sets such that $\bigcap_\alpha S_\alpha \neq \emptyset$. Then $\bigcap_\alpha S_\alpha \in P$.*

Lemma 2. *Let $S \in P$ and $\Lambda \in X^*$. Then the set $S_\Lambda = \{x \in S : \Lambda x = \max_S \Lambda\}$ belongs to P.*

We claim that every $S \in P$ contains an extreme point. To prove this, set

$$P_S = \big\{A \in P : A \subset S\big\}.$$

We define a partial order on P_S as follows: $S_1 \leqslant S_2$ if and only if $S_1 \supset S_2$.

(a) Prove (with the help of Lemma 1) that P_S is inductive, and apply Zorn's lemma to deduce that P_S admits a maximal element M.

(b) Invoke the separation theorem to prove (with the help of Lemma 2) that M is a singleton $\{x\}$.

(c) Show that x is an extreme point of S, which establishes the claim.

[1] The proof follows Rudin [38, Theorem 3.21].

The remainder of the proof goes as follows. Let E be the (nonempty) set of extreme points of K, and set $H = \operatorname{co} E$. The arguments above show that $H \cap S \neq \emptyset$ for every extreme set $S \in P$. We have $\operatorname{cl} H \subset K$, and so $\operatorname{cl} H$ is compact. We now conclude by showing that $\operatorname{cl} H = K$, arguing by contradiction.

Suppose there is a point $\bar{x} \in K \backslash \operatorname{cl} H$. By the separation theorem, there exist $\Lambda \in X^*$ and γ such that

$$\Lambda x < \gamma < \Lambda \bar{x} \quad \forall x \in \operatorname{cl} H.$$

This implies $\operatorname{cl} H \cap K_\Lambda = \emptyset$, where

$$K_\Lambda = \{x \in K : \Lambda x = \max_K \Lambda\}.$$

But $K_\Lambda \in P$ by Lemma 2; this is the required contradiction. □

The reader should note a corollary of the theorem:

A nonempty compact convex subset of a normed space admits an extreme point.

8.57 Exercise. The (strong) compactness of K in the theorem above is at times a restrictive hypothesis. Show that the proof can be adapted to the two following cases, in which compactness is provided by weak topologies:

(a) X is a reflexive Banach space, and K is closed, convex, and bounded.

(b) X is the dual of a normed space Y, and K is the unit ball in X. □

8.58 Exercise. For $1 \leqslant p \leqslant \infty$, let X be the Banach space $L^p(0,1)$, and B its closed unit ball.

(a) If $p = 1$, show that B has no extreme points.

(b) Deduce from Exer. 8.57 that $L^1(0,1)$ is not isometric to the dual of a normed space (which implies that it is not reflexive).

(c) If $1 < p < \infty$, show that every point in the unit sphere is an extreme point of B.

(d) Determine the extreme points of B in the case $p = \infty$. □

Part II
Optimization and Nonsmooth Analysis

Chapter 9

Optimization and multipliers

I think the restriction to smooth manifolds is dictated only by technical reasons and is unsatisfactory...non-smooth manifolds exist naturally.

Shiing-Shen Chern

Consider God's handiwork: who can straighten what He hath made crooked?

Ecclesiastes 7:13

The abstract optimization problem $\min_A f$ consists of minimizing a *cost function* $f(x)$ over the points x belonging to the *admissible set* A. The set A incorporates the constraints imposed upon the points that are allowed to compete in the minimization. The nature of A, and also of the function f, determine whether our problem is classical or modern, discrete or continuous, finite or infinite dimensional, smooth or convex.

Optimization is a rich and varied subject with numerous applications. The core mathematical issues, however, are always the same:

- Existence: Is there, in fact, a solution of the problem? (This means a point $x_* \in A$ at which $\min_A f$ is attained.)
- Necessary conditions: What special properties must a solution have, properties that will help us to identify it?
- Sufficient conditions: Having identified a point that is suspected of being a solution, what tools can we apply to confirm the suspicion?

Many other issues than these can, and do arise, depending on the nature of the problem. Consider for example the calculus of variations, which we take up later on, in which the variable x refers to a function. The regularity of the minimizing function x_* reveals itself to be a central question, one that we shall study rather thoroughly. In contrast, issues such as modeling, computation, and implementation, which are crucial to applied optimization of all sorts, are not on our agenda.

Deductive and inductive methods. A familiar optimization problem that the reader has encountered is that of minimizing a differentiable function $f : \mathbb{R}^n \to \mathbb{R}$ over all points x belonging to $A = \mathbb{R}^n$. This is a "free" optimization problem, since the admissible points x are subject to no explicit constraint (except, of course, that they reside in the underlying space $\mathbb{R}^n$).

F. Clarke, *Functional Analysis, Calculus of Variations and Optimal Control*, Graduate Texts in Mathematics 264, DOI 10.1007/978-1-4471-4820-3_9,

The existence question might be treated by imposing a supplementary hypothesis on f, for example, a growth condition: $f(x) \to +\infty$ as $|x| \to +\infty$. Then, bearing in mind that f is continuous (since it is differentiable), it follows that a solution x_* does exist. We would like to identify it, however.

For that purpose, we turn to necessary conditions, proceeding to invoke *Fermat's rule*: a solution x_* must satisfy $\nabla f(x_*) = 0$. (It is in order to write this equation that f was taken to be differentiable, rather than merely continuous.) Thus, we search for a solution among the critical points of f.

Then, we could conclude in one of two ways. If we know that a solution exists, and if we have examined the critical points in order to find the best critical point x_* (that is, the one assigning the lowest value to the cost f), it follows logically that x_* is the solution to the minimization problem. This approach is known as the **deductive method**.[1]

There is a potential fallacy lurking here, one that is rather common in certain areas of application. It consists of applying the deductive reasoning above without knowing with certainty that a solution exists. In the absence of guaranteed existence, it is quite possible for the necessary conditions to identify a unique admissible point x, which then fails to be a solution (because there isn't one).

An alternative approach, one that does not necessitate finding all the critical points, or knowing ahead of time that a solution exists, is to find an argument tailored precisely to a given suspect x_*. Let us give three examples of how this might work. First, suppose we find it possible to rewrite $f(x)$ as follows:

$$f(x) = \big[\varphi(x) - \varphi(x_*)\big]^2 + c,$$

for some function φ and constant c. Then, evidently, x_* minimizes f.

Another strategy would be to postulate the convexity of the function f; then, the stationarity condition $\nabla f(x_*) = 0$ implies, without further argument, that x_* is a global minimum for f.

Finally, let us mention a third tactic: if f is twice continuously differentiable, the condition $\nabla^2 f(x_*) > 0$ (positive definite) together with $\nabla f(x_*) = 0$, is enough to imply that f is at least a local minimum.

Note that all three of these alternate arguments do *not* require an existence theorem. They are examples of the **inductive method**.[2] These two approaches to solving optimization problems, the deductive and the inductive, will play a role in shaping the things to come.

We turn now to the issue of necessary conditions in the presence of constraints.

[1] Deductive: reasoning from the general to the particular.

[2] Inductive reasoning: wherein one argues from the particular to the general.

9.1 The multiplier rule

Consider the following special case of our problem $\min_A f$:

$$\text{Minimize } f(x) \text{ subject to } h(x) = 0. \tag{P$_0$}$$

Here, the admissible set is given by $A = \{x \in \mathbb{R}^n : h(x) = 0\}$, and we are dealing with a *constrained optimization* problem, one in which admissibility is defined via an equality constraint. To keep things simple, we suppose for now that f and h are continuously differentiable, and that h is real-valued.

There is a famous technique for obtaining necessary conditions in this case, known as *Lagrange multipliers*. It should be part of any mathematical education, for it is a serious nominee for the most useful theorem in applied mathematics.

The method consists of seeking the solutions of the constrained problem (P$_0$) above among the critical points, not of f (for this would ignore the constraint), but of $f + \lambda h$, where the *multiplier* λ is a parameter whose value is not known for the moment. The resulting equation $\nabla(f + \lambda h)(x) = 0$ may appear to be a step in the wrong direction, since it involves an additional unknown λ, but this is compensated for by the constraint equation $h(x) = 0$. The idea is to solve the two equations for x and λ simultaneously, and thus identify x (and λ, for whatever that's worth).

The theorem we have alluded to is known as the *multiplier rule*. We now discuss in some detail (but in general terms) how to prove such necessary conditions for optimality, as they are known.

Terminology: Various branches of optimization employ different synonyms for a "solution" of the underlying problem. A point x_* that solves the minimization problem can be called *optimal*, or it can be referred to as a *minimizer*, or it can be said that it provides a minimum. The word "local" is used in addition, when the minimum in question is a local one in some prescribed sense.

The first approach to proving the multiplier rule is geometric. Let x_* solve (P$_0$), and consider, for $\varepsilon > 0$, the relation of the set

$$f_\varepsilon^{-1} := \{x \in \mathbb{R}^n : f(x) = f(x_*) - \varepsilon\}$$

to the admissible set $A = \{x \in \mathbb{R}^n : h(x) = 0\}$. Clearly, these two sets (which we imagine as surfaces in $\mathbb{R}^n$) do not intersect, for otherwise x_* cannot be a solution of (P$_0$). As ε decreases to 0, the surfaces f_ε^{-1} "converge" to the level set

$$\{x \in \mathbb{R}^n : f(x) = f(x_*)\},$$

which *does* have a point in common with A, namely x_*. (Have we mentioned that we are arguing in general terms?) Thus, the value $\varepsilon = 0$ corresponds to a point of first contact (or "osculation") between these surfaces. We conclude that the normal

vectors at x_* of these two surfaces are parallel, that is, multiples of one another. Since normals to level sets are generated by gradients, we deduce the existence of a scalar λ such that $\nabla f(x_*) + \lambda \nabla h(x_*) = 0$. This is precisely the multiplier rule we seek to establish.

The argument given above has the merit of explaining the geometric meaning behind the multiplier rule. It is difficult to make it rigorous, however. A more manageable classical approach is to consider the nature of the solutions $(x,r) \in \mathbb{R}^n \times \mathbb{R}$ of the equation

$$F(x,r) := \big(f(x) - f(x_*) + r,\, h(x)\big) = (0,0).$$

The reader will observe that the point $(x_*, 0)$ satisfies the equation.

If the Jacobian matrix $D_x F(x_*, 0)$ has (maximal) rank 2, then, by the implicit function theorem, the equation $F(x,r) = (0,0)$ admits a solution $x(r)$ for every r near 0, where $\lim_{r \to 0} x(r) = x_*$. But then, for $r > 0$ sufficiently small, we obtain a point $x(r)$ arbitrarily near x_* which is admissible, and for which $f(x(r)) < f(x_*)$. This contradicts even the local optimality of x_*. It follows that the rows of $D_x F(x_*, 0)$, namely the vectors $\nabla f(x_*)$ and $\nabla h(x_*)$ (modulo transpose), must be linearly dependent. If we assume that $\nabla h(x_*) \neq 0$ (as is usually done), this implies that, for some λ, we have $\nabla f(x_*) + \lambda \nabla h(x_*) = 0$. Ergo, the multiplier rule.

This classical argument is satisfyingly rigorous, but it is difficult to adapt it to different types of constraints, notably inequality constraints $g(x) \leqslant 0$, and unilateral constraints $x \in S$. Other considerations, such as replacing $\mathbb{R}^n$ by an infinite dimensional space, or allowing the underlying functions to be nondifferentiable, further complicate matters.

Let us turn, then, to an entirely different argument for proving the multiplier rule, one that we invite the reader to criticize. It is based upon considering the following *perturbed* problem (P_α):

$$\text{Minimize } f(x) \text{ subject to } h(x) = \alpha. \qquad (P_\alpha)$$

Note that the original problem (P_0) has been imbedded in a family of problems depending on the parameter α. We define $V(\alpha)$ to be the value of the minimum in the problem (P_α). Thus, by definition of V, and since x_* solves (P_0) by assumption, we have $V(0) = f(x_*)$. On the other hand, for any x, the very definition of V implies that $V(h(x)) \leqslant f(x)$. (There is a pause here while the reader checks this.) We may summarize these two observations as follows:

$$f(x) - V\big(h(x)\big) \geqslant 0 \;\; \forall x, \text{ with equality for } x = x_*.$$

By Fermat's rule, the gradient of the function in question must vanish at x_*. By the chain rule, we obtain:

$$\nabla f(x_*) - V'\big(h(x_*)\big) \nabla h(x_*) = \nabla f(x_*) - V'(0) \nabla h(x_*) = 0.$$

Behold, once again, the multiplier rule, with $\lambda = -V'(0)$. We have also gained new insight into the meaning of the multiplier λ: it measures the sensitivity of the problem with respect to perturbing the equality constraint $h = 0$ to $h = \alpha$. (This interpretation is well known in such fields as operations research, mechanics, or economics.)

Nonsmoothness. The "value function proof" that we have just presented is completely rigorous, if it so happens that V is differentiable at 0. It must be said at once, however, that value functions are notoriously nonsmooth. Note that V above is not even finite-valued, necessarily: $V(\alpha) = +\infty$ when the set $\{x : h(x) = \alpha\}$ is empty. And simple examples show that V is not necessarily differentiable, even when it is finite everywhere. This raises the issue of rescuing the proof through the use of generalized derivatives and nonsmooth calculus, subjects that we develop in subsequent chapters.

Let us mention one more approach to proving the multiplier rule, one that uses an important technique in optimization: *exact penalization*. Our interest remains focused on the problem $\min_A f$, but we consider the (free!) minimization of the function $f(x) + k\,d_A(x)$, where k is a positive number and, as usual, d_A denotes the distance function associated with A. Under mild hypotheses, it turns out that for k sufficiently large, the solution x_* of the constrained problem will be a local solution of this unconstrained problem. We might say that the constraint has been absorbed into the cost by penalization.

At this point, we are tempted to write Fermat's rule: $\nabla(f + k\,d_A)(x_*) = 0$. There is a difficulty, once more having to do with regularity, in doing so: distance functions like d_A are not differentiable. Once again, then, we require some generalized calculus in order to proceed. A further issue also arises: given that A is the set $\{h = 0\}$, how may we interpret the generalized derivative of d_A? Is it characterized by ∇h somehow, and would this lead (yet again) to the multiplier rule? We shall develop later the "nonsmooth geometry" required to answer such questions (positively).

We have explained how considerations of theory lead to nonsmoothness. In fact, there are many important problems that feature data that are nondifferentiable from the start. They arise in such areas as elasticity and mechanics, shape optimization and optimal design, operations research, and principal-agent analysis in economics. However, we begin our study with the smooth case, in a more general setting as regards the constraints that define admissibility.

The basic problem. The focus of this chapter is the following basic problem of constrained optimization:

$$\text{Minimize } f(x) \text{ subject to } g(x) \leqslant 0,\ \ h(x) = 0,\ \ x \in S \qquad \text{(P)}$$

where the functions

$$f : X \to \mathbb{R}, \quad g : X \to \mathbb{R}^m, \quad h : X \to \mathbb{R}^n,$$

together with the set S in the Banach space X, constitute the given data. The vector inequality $g(x) \leqslant 0$ means, of course, that each component $g_i(x)$ of $g(x)$ satisfies $g_i(x) \leqslant 0$ $(i = 1, 2, \ldots, m)$. An optimization problem of this type is sometimes referred to as a *program*, which is why the term *mathematical programming* is a synonym for certain kinds of optimization.

Terminology. We say that $x \in X$ is *admissible* for the problem (P) if it lies in S and satisfies both the inequality constraint $g(x) \leqslant 0$ and the equality constraint $h(x) = 0$. The requirement $x \in S$ is also referred to as the *unilateral constraint*. A *solution* x_* of (P) is an admissible point which satisfies $f(x_*) \leqslant f(x)$ for all other admissible points x, where f is the *cost function*. We also say that x_* is *optimal* for the problem, or is a *minimizer*.

9.1 Theorem. (Multiplier rule) *Let x_* be a solution of* (P) *that lies in the interior of S. Suppose that all the functions involved are continuously differentiable in a neighborhood of x_*. Then there exists $(\eta, \gamma, \lambda) \in \mathbb{R} \times \mathbb{R}^m \times \mathbb{R}^n$ satisfying the* **nontriviality** *condition*

$$(\eta, \gamma, \lambda) \neq 0,$$

together with the **positivity** *and* **complementary slackness** *conditions*

$$\eta = 0 \text{ or } 1, \ \ \gamma \geqslant 0, \ \ \langle \gamma, g(x_*) \rangle = 0,$$

and the **stationarity** *condition*

$$\big\{ \eta f + \langle \gamma, g \rangle + \langle \lambda, h \rangle \big\}'(x_*) = 0.$$

Remarks on the multiplier rule. The triple (η, γ, λ) is called a **multiplier**. The theorem asserts that the existence of such a multiplier is a necessary condition for x_* to be a solution. The term *Lagrange multiplier* is often used, in honor of the person who used the concept to great effect in classical mechanics; in fact, the idea goes back to Euler (1744). The hypothesis that the functions involved are smooth, and that x_* lies in the interior of S, makes the setting of the theorem a rather classical one, though the combination of an infinite dimensional underlying space with the presence of mixed equality/inequality constraints is modern.

Because we have assumed $x_* \in \operatorname{int} S$, the set S plays no role in the necessary conditions. In fact, S merely serves (for the present) to localize the optimization problem. Suppose, for example, that x_* is merely a *local* minimum for (P), when the unilateral constraint $x \in S$ is absent. Then, by adding the constraint $x \in S$, where S is a sufficiently small neighborhood of x_*, we transform x_* into a *global* minimum. Another possible role of S is to define a neighborhood of x_* in which certain hypotheses hold (in this case, continuous differentiability).

The reader will observe that the nontriviality condition is an essential component of the multiplier rule, since the theorem is vacuous in its absence: the triple $(0,0,0)$ satisfies all the other conclusions, for any x_*.

The complementary slackness condition $\langle \gamma, g(x_*) \rangle = 0$ is equivalent to

$$i \in \{1,2,\ldots,m\},\ \ g_i(x_*) < 0 \implies \gamma_i = 0.$$

This equivalence follows from the observation that, since $\gamma \geqslant 0$ and $g(x_*) \leqslant 0$, the inner product $\langle \gamma, g(x_*) \rangle$ is necessarily nonpositive, and equals zero if and only if each term $\gamma_i\, g_i(x_*)$ is zero. Thus, we may rephrase the complementary slackness condition as follows: if the constraint $g_i \leqslant 0$ is not *saturated* at x_* (that is, if $g_i(x_*) < 0$), then the function g_i does not appear in the necessary conditions (the corresponding γ_i is equal to 0). This makes perfect sense, for if $g_i(x_*) < 0$, then (by the continuity of g) we have $g_i(x) < 0$ for all nearby points x, so that (locally) the constraint is redundant, and can be ignored.

The case $\eta = 0$ of the multiplier rule yields necessary conditions that do not involve the cost function f. Typically, this rather pathological situation arises when the equality and inequality constraints are so “tight” that they are satisfied by just one point x_*, which is then *de facto* optimal, independently of f. The case $\eta = 0$ is referred to as the **abnormal** case. In contrast, when $\eta = 1$, we say that we are in the **normal** case.

The proof of Theorem 9.1 is postponed to §10.4, where, using techniques of nonsmooth analysis, a more general result can be proved.

Absence of certain constraints. Either equality or inequality constraints can be absent in the problem treated by Theorem 9.1, which then holds without reference to the missing data. Consider first the case of the problem in which there are no inequality constraints. We can simply introduce a function g that is identically -1, and then apply the theorem. When we examine the resulting necessary conditions, we see that the multiplier γ corresponding to g must be 0, and therefore the conclusions can be couched entirely in terms of a nontrivial multiplier (η, λ), with no reference to g. Note that the resulting multiplier must be normal if the vectors $h_j'(x_*)$ $(j = 1,2,\ldots,n)$ are independent. This assumption, a common one, is said to correspond to the nondegeneracy of the equality constraints.

Consider next the problem having no equality constraints. Let us introduce another variable $y \in \mathbb{R}$ on which f and g have no dependence. In $X \times \mathbb{R}$, we redefine S to be $S \times \mathbb{R}$, and we impose the equality constraint $h(y) := y = 0$. We then proceed to apply Theorem 9.1 to this augmented problem. There results a multiplier (η, γ, λ); the stationarity with respect to y yields $\lambda = 0$. Then all the conclusions of the theorem hold for a nontrivial multiplier (η, γ) (with no reference to h and λ).

The meaning of the multiplier rule. Consider the problem (P) in the case when only inequality constraints are present. For any admissible x, we denote by $I(x)$ the set of indices for which the corresponding inequality constraint is active at x:

$$I(x) = \big\{ i \in \{1,2,\ldots,m\} : g_i(x) = 0 \big\}.$$

Now let x_* be optimal. As a consequence of this optimality, we claim that there cannot exist $v \in X$ such that, simultaneously,

$$\langle f'(x_*), v \rangle < 0 \text{ and } \langle g_i'(x_*), v \rangle < 0 \;\; \forall i \in I(x_*).$$

Such a v would provide a direction in which, for a small variation, the function f, as well as each g_i for $i \in I(x_*)$, would simultaneously decrease. Thus, for all $t > 0$ sufficiently small, we would have

$$f(x_* + tv) < f(x_*), \;\; g_i(x_* + tv) < g_i(x_*) = 0 \;\; \forall i \in I(x_*).$$

But then, by further reducing t if necessary, we could arrange to have

$$g_i(x_* + tv) \leqslant 0 \text{ for \emph{all} indices } i \in \{1, 2, \ldots, m\},$$

as well as $f(x_* + tv) < f(x_*)$. This would contradict the optimality of x_*.

The nonexistence of such a direction v is equivalent to the positive linear dependence of the set

$$\{ f'(x_*), \; g_i'(x_*) : i \in I(x_*) \},$$

as Exer. 2.40 points out. We conclude, therefore, that the necessary conditions of the multiplier rule correspond to the nonexistence of a decrease direction (in the above sense). (In fact, this is a common feature of first-order necessary conditions in various contexts.) We remark that in the presence of equality constraints, it is much harder to argue along these lines, especially in infinite dimensions.

9.2 Exercise. We wish to minimize $f(x)$ subject to the constraints $g_1(x) \leqslant 0$ and $g_2(x) \leqslant 0$, where f, g_1 and g_2 are continuously differentiable functions defined on $\mathbb{R}^3$. At four given points x_i in $\mathbb{R}^3$ $(i = 1, 2, 3, 4)$, we have the following data:

	g_1	g_2	∇f	∇g_1	∇g_2
x_1	0	0	$(2,-2,4)$	$(-2,0,0)$	$(0,1,-2)$
x_2	0	-1	$(0,1,1)$	$(0,-1,0)$	$(0,0,-1)$
x_3	0	1	$(0,0,1)$	$(0,0,-1)$	$(0,0,-1)$
x_4	0	0	$(1,1,1)$	$(0,-1,0)$	$(1,0,1)$

(a) Only one of these four points could solve the problem. Which one is it?

(b) For each point x_i that is admissible but definitely not optimal, find a direction in which a small displacement can be made so as to attain a “better” admissible point. □

9.3 Example. We allow ourselves to hope that the reader has seen the multiplier rule applied before. However, just in case the reader’s education has not included this topic, we consider now a simple ‘toy problem’ (to borrow a phrase from the

physicists) of the type that the author seems to recall having seen in high school (but that was long ago). Despite its simplicity, some useful insights emerge.

The problem is that of designing a soup can of maximal volume V, given the area q of tin that is available for its manufacture. It is required, for reasons of solidity, that the thickness of the base and of the top must be double that of the sides. More specifically then, we wish to find the radius x and the height y of a cylinder such that the volume $V = \pi x^2 y$ is maximal, under the constraint $2\pi xy + 4\pi x^2 = q$. We do not doubt the reader's ability to solve this problem without recourse to multipliers, but let us do so by applying Theorem 9.1.

We could view the constraint $2\pi xy + 4\pi x^2 = q$ as an equality constraint (which it is), but we can also choose to replace it by the inequality constraint

$$g(x,y) := 2\pi xy + 4\pi x^2 - q \leqslant 0,$$

since it is clear that the solution will use all the available tin. Doing so offers the advantage of knowing beforehand the sign of the multiplier that will appear in the necessary conditions.

The problem has a natural (implicit) constraint that x and y must be nonnegative, a feature of many optimization problems that motivates the following definition.

Notation. We denote by $\mathbb{R}^n_+$ the set $\{x \in \mathbb{R}^n : x \geqslant 0\}$, also referred to as the positive orthant.

To summarize, then, we have the case of problem (P) in which $(x,y) \in \mathbb{R}^2$ and

$$f(x,y) := -\pi x^2 y, \quad g(x,y) = 2\pi xy + 4\pi x^2 - q, \quad S = \mathbb{R}^2_+$$

with the equality constraint $h = 0$ being absent. (Note the minus sign in f, reflecting the fact that our theory was developed for minimization rather than maximization.) If we take q strictly positive, it is easy to prove that a solution (x_*, y_*) of the problem exists, and that we have $x_* > 0,\ y_* > 0$ (thus, the solution lies in $\operatorname{int} S$).

The usual first step in applying Theorem 9.1 is to rule out the abnormal case $\eta = 0$; we proceed to do so. If $\eta = 0$, then the necessary conditions imply that $\nabla g(x_*, y_*)$ equals $(0,0)$, which leads to $x_* = y_* = 0$, which is absurd. Thus, we may take $\eta = 1$. (Note that the abnormal case corresponds to an exceedingly tight constraint, the case $q = 0$.) With $\eta = 1$, the resulting stationarity condition becomes

$$-2\pi x_* y_* + \gamma\,(2\pi y_* + 8\pi x_*) = 0, \quad -\pi x_*^2 + \gamma\,(2\pi x_*) = 0.$$

The second equation gives $x_* = 2\gamma$, whence $\gamma > 0$; substituting in the first equation then produces $y_* = 8\gamma$. Since $\gamma > 0$, the inequality constraint is saturated (as expected). The equality $g(x_*, y_*) = 0$ then leads to

$$\gamma = \sqrt{q}/\big(4\sqrt{3\pi}\,\big), \quad y_* = 4x_*, \quad f(x_*, y_*) = -q^{3/2}/\big(6\sqrt{3\pi}\,\big).$$

Thus the height of the optimal soup can is four times its radius. As regards our own (non soup oriented) intentions, it is more to the point to note that the derivative of the optimal volume $q^{3/2}/(6\sqrt{3\pi}\,)$ with respect to q is precisely the multiplier γ, thus confirming the interpretation of γ (suggested in the introduction) as a sensitivity with respect to changing the constraint (that is, the amount of available tin). Economists would refer to γ as a *shadow price*.[3] □

9.2 The convex case

The next item on the agenda is to impart to the reader an appreciation of the "convex case" of the problem (P). We shall see in this section that the multiplier rule holds in a stronger form in this setting, and that it is normally sufficient as well as necessary. In the following section, another characteristic feature of convex optimization is examined: the possibility of defining a useful *dual problem*. Together, these elements explain why, other things being equal, the convex case of (P) is preferred, if we can so arrange things.

The problem (P) is unaltered: it remains that of minimizing $f(x)$ subject to the constraints

$$g(x) \leqslant 0\,,\ \ h(x) = 0\,,\ \ x \in S\,,$$

but in the following framework, referred to as the *convex case*:

- S is a convex subset of a real vector space X;
- The following functions are convex:

$$f : S \to \mathbb{R}\ \text{ and }\ g_i : S \to \mathbb{R}\ (i = 1, 2, \ldots, m);$$

- Each function $h_j : S \to \mathbb{R}\ (j = 1, 2, \ldots, n)$ is affine; that is, h_j is of the form $\langle\, \zeta_j, x\rangle + c_j$, where ζ_j is a linear functional on S and $c_j \in \mathbb{R}$.

Note that these functions need only be defined on S. In the following counterpart to Theorem 9.1, it is *not* required that x_* lie in the interior of S; indeed, no topology is imposed on X.

9.4 Theorem. (Kuhn-Tucker) *Let x_* be a solution of* (P) *in the convex case. Then there exists $(\eta, \gamma, \lambda) \in \mathbb{R} \times \mathbb{R}^m \times \mathbb{R}^n$ satisfying the* **nontriviality** *condition*

$$(\eta, \gamma, \lambda) \neq 0\,,$$

[3] The shadow price would be used, for example, to decide whether the soup can (which is optimal for the specified volume) should be made larger (in order to increase profit). To decide, one compares $p\,\gamma$ (the marginal effect on revenue of using more tin, where p is the unit price of soup) to the marginal cost of tin; at optimality, the two marginal effects are equal.

the **positivity** *and* **complementary slackness** *conditions*

$$\eta = 0 \text{ or } 1\,,\;\; \gamma \geqslant 0\,,\;\; \langle \gamma, g(x_*) \rangle = 0\,,$$

and the **minimization** *condition*

$$\big\{ \eta f + \langle \gamma, g \rangle + \langle \lambda, h \rangle \big\}(x) \;\geqslant\; \big\{ \eta f + \langle \gamma, g \rangle + \langle \lambda, h \rangle \big\}(x_*) \;=\; \eta f(x_*) \;\;\forall x \in S\,.$$

Proof. We consider the following subset of $\mathbb{R} \times \mathbb{R}^m \times \mathbb{R}^n$:

$$C \;=\; \big\{ \big(f(x) + \delta,\, g(x) + \Delta,\, h(x) \big) \;:\; \delta \geqslant 0,\; \Delta \geqslant 0,\; x \in S \big\}\,.$$

It is easy to see that C is convex (that's what the convexity hypotheses on the data are for). We claim that the point $(f(x_*), 0, 0)$ lies in the boundary of C.

If this were false, C would contain, for some $\varepsilon > 0$, a point of the form

$$\big(f(x) + \delta,\, g(x) + \Delta,\, h(x) \big) \;=\; (f(x_*) - \varepsilon, 0, 0)\,, \text{ where } x \in S\,,\; \delta \geqslant 0\,,\; \Delta \geqslant 0\,.$$

But then x is admissible for (P) and assigns to f a strictly lower value than does x_*, contradicting the optimality of x_*.

Since C is finite dimensional, the normal cone (in the sense of convex analysis) to C at this boundary point is nontrivial (Cor. 2.48). This amounts to saying that there exists $(\eta, \gamma, \lambda) \neq 0$ such that

$$\eta\big(f(x) + \delta\big) + \langle \gamma, g(x) + \Delta \rangle + \langle \lambda, h(x) \rangle \;\geqslant\; \eta f(x_*) \;\;\forall x \in S\,,\; \delta \geqslant 0\,,\; \Delta \geqslant 0\,.$$

Note that this yields the minimization condition of the theorem. It also follows readily that $\eta \geqslant 0$ and $\gamma \geqslant 0$. Taking

$$x = x_*\,,\;\; \delta = 0\,,\;\; \Delta = 0$$

in the inequality gives $\langle \gamma, g(x_*) \rangle \geqslant 0$, which is equivalent to the complementary slackness condition $\langle \gamma, g(x_*) \rangle = 0$, since $g(x_*) \leqslant 0$ and $\gamma \geqslant 0$. Finally, if $\eta > 0$, note that we can normalize the multiplier (η, γ, λ); that is, replace it by $(1, \gamma/\eta\,, \lambda/\eta)$. Thus, in all cases, we can assert that η equals 0 or 1. □

Remark. We refer to the vector (η, γ, λ) as a *multiplier in the convex sense*. The difference between such a multiplier and a classical one (as given in Theorem 9.1) is that the stationarity is replaced by an actual minimization. Furthermore, no differentiability of the data is assumed here, and, as we have said, there is no requirement that x_* lie in the interior of S.

In the same vein as our discussion following Theorem 9.1, it is easy to see that the theorem above adapts to the cases in which either the equality or inequality constraint is absent, by simply deleting all reference to the missing constraint.

9.5 Example. (du Bois-Raymond lemma) The following situation arises in the calculus of variations, where the conclusion below will be of use later on. We are given an element $\theta \in L^1(a,b)$ such that

$$\int_a^b \theta(t)\,\varphi'(t)\,dt \;=\; 0 \;\;\forall\, \varphi \in \operatorname{Lip}_0[a,b],$$

where $\operatorname{Lip}_0[a,b]$ is the set of Lipschitz functions on $[a,b]$ that vanish at a and b. Evidently, the stated condition holds if θ is constant; our goal is to prove that this is the only case in which it holds.

Let X be the vector space of all $\varphi \in \operatorname{Lip}[a,b]$ satisfying $\varphi(a) = 0$, and define

$$f(\varphi) \;=\; \int_a^b \theta(t)\,\varphi'(t)\,dt\,, \quad h(\varphi) \;=\; \varphi(b)\,.$$

Then, by hypothesis, we have $f(\varphi) \geqslant 0$ for all $\varphi \in X$ satisfying $h(\varphi) = 0$. Thus the function $\varphi_* \equiv 0$ solves the corresponding version of the optimization problem (P). We proceed to apply Theorem 9.4. Accordingly, there exists a multiplier $(\eta, \lambda) \neq 0$ with η equal to 0 or 1, such that

$$\eta \int_a^b \theta(t)\,\varphi'(t)\,dt + \lambda\,\varphi(b) \;=\; \int_a^b \{\,\eta\,\theta(t) + \lambda\,\}\,\varphi'(t)\,dt \;\geqslant\; 0 \;\;\forall\, \varphi \in X\,.$$

It follows that $\eta = 0$ cannot occur, for then we would have $\lambda = 0$ too, violating nontriviality. Thus we may set $\eta = 1$, and we obtain

$$\int_a^b \{\,\theta(t) + \lambda\,\}\,\varphi'(t)\,dt \;\geqslant\; 0 \;\;\forall\, \varphi \in X\,.$$

For a positive integer k, let A_k be the set $\{t \in (a,b) : |\theta(t)| \leqslant k\}$, and let χ_k be its characteristic function. Taking

$$\varphi(t) \;=\; -\int_a^t \{\,\theta(s) + \lambda\,\}\,\chi_k(s)\,ds$$

in the inequality above (note that $\varphi \in X$) yields

$$-\int_{A_k} \{\,\theta(t) + \lambda\,\}^2\,dt \;\geqslant\; 0\,.$$

Thus the integral is 0 for every k, and we discover $\theta(t) + \lambda \;=\; 0$ a.e. ☐

9.6 Exercise. Let x_* be a solution of the problem encountered in Exer. 5.53. Show that the problem fits into the framework of Theorem 9.4 if one takes

$$g(x) \;=\; \sum_{i=1}^{\infty} x_i - 1,\;\; S = \Big\{\, x \in \ell^r : 0 \leqslant x_i \,\forall\, i,\; \sum_{i=1}^{\infty} x_i < \infty,\; f(x) < \infty \Big\}$$

and if no equality constraint is imposed. Deduce the existence of a nonnegative constant γ such that, for each i, the value x_{*i} minimizes the function $t \mapsto f_i(t)+\gamma t$ over $[0,\infty)$. In that sense, all the x_{*i} are determined by a single scalar γ. □

Remark. In many cases, the minimization condition in the theorem can be expressed in equivalent terms as a stationarity condition, via the subdifferential and the normal cone of convex analysis. This brings out more clearly the common aspects of Theorem 9.1 and Theorem 9.4, as we now see.

9.7 Exercise.

(a) Let x_* be admissible for (P) in the convex case, and suppose there exists a normal multiplier $(1,\gamma,\lambda)$ associated with x_*. Prove that x_* is optimal.

(b) In addition to the hypotheses of Theorem 9.4, suppose that X is a normed space, and that f, g, and h are convex and continuous on X. Prove that the minimization condition in the conclusion of the theorem is equivalent to

$$0 \in \partial\big\{\eta f+\langle\gamma,g\rangle+\langle\lambda,h\rangle\big\}(x_*)+N_S(x_*)\,.$$

Under what additional hypotheses would this be equivalent to the stationarity conclusion in Theorem 9.1? □

The exercise above expresses the fact that in the convex case of (P), the necessary conditions, when they hold normally, are also sufficient for optimality. Another positive feature of the convex case is the possibility of identifying reasonable conditions in the presence of which, *a priori*, the necessary conditions must hold in normal form. We are referring to the **Slater condition**, which is said to hold when:

- X is a normed space;
- There exists a *strictly admissible* point x_0 for (P):

$$x_0 \in \operatorname{int} S\,, \quad g(x_0) < 0\,, \quad h(x_0) = 0\,;$$

- The affine functions of the equality constraint are independent, meaning that the set $\{h'_j \,:\, j=1,2,\ldots,n\}$ is independent.

9.8 Theorem. *In the convex case of problem* (P), *when the Slater condition holds, the multiplier whose existence is asserted by* Theorem 9.4 *is necessarily normal*: $\eta = 1$.

Proof. We reason *ad absurdum*, by supposing that $(0,\gamma,\lambda)$ is an abnormal (nontrivial) multiplier. The minimization condition, when expressed at the point x_0 provided by the Slater condition, gives $\langle\gamma, g(x_0)\rangle \geqslant 0$. Since every component of $g(x_0)$ is strictly negative, and since $\gamma \geqslant 0$, we deduce $\gamma = 0$. Then the minimization condition becomes: $\langle\lambda, h(x)\rangle \geqslant 0\ \ \forall x \in S$. Since equality holds at $x_0 \in \operatorname{int} S$, we have

$\sum_i \lambda_i h_i' = 0$ by Fermat's rule. Then the linear independence implies $\lambda = 0$, contradicting the nontriviality of the multiplier. □

We now illustrate the use of the Slater condition by means of a simple problem arising in statistics.

9.9 Exercise. Let $z_1, z_2, \ldots, z_n$ be the n distinct values of a random variable Z, and let p_i be the probability that $Z = z_i$. Let us suppose that we know from observation that Z has mean value m, so that $\sum_i z_i p_i = m$. However, the probabilities p_i are not known. A common way to estimate the probability distribution $p = (p_1, p_2, \ldots, p_n)$ in this case is to postulate that it maximizes the *entropy*

$$E = -\sum_{i=1}^{n} p_i \ln p_i .$$

The optimization problem, then, is to maximize E subject to the constraints

$$p \in \mathbb{R}^n_+ , \quad \sum_{i=1}^{n} p_i = 1 , \quad \sum_{i=1}^{n} z_i p_i = m .$$

We place this in the context of Theorem 9.4 by taking $X = \mathbb{R}^n$ and $S = \mathbb{R}^n_+$, and by defining

$$f(p) = \sum_i p_i \ln p_i , \quad h_1(p) = \big(\sum_i p_i \big) - 1 , \quad h_2(p) = \big(\sum_i z_i p_i \big) - m .$$

Thus, the equality constraint has two components, and the inequality constraint is absent. Note that the function $t \mapsto t \ln t$ has a natural value of 0 at $t = 0$.

(a) Prove that f is convex on S.

(b) Prove that a solution to the problem exists, and that it is unique.

We suppose henceforth that $\min_i z_i < m < \max_i z_i$. If this were not the case, m would equal either $\min_i z_i$ or $\max_i z_i$, which means that the distribution has all its mass on a single value: a case of overly tight constraints which, of themselves, identify the solution.

(c) Prove that the Slater condition is satisfied. Deduce that the solution admits a normal multiplier in the convex sense.

(d) Deduce from the minimization condition of the multiplier that the solution p satisfies $p_i > 0 \ \forall i \in \{1, 2, \ldots, n\}$.

(e) Prove that the solution p corresponds to an *exponential* distribution: for certain constants c and k, we have

$$p_i = \exp\big(c + k z_i\big) , \quad i = 1, 2, \ldots, n . \qquad \square$$

The next result gives a precise meaning (in the current convex setting) to the interpretation of multipliers in terms of sensitivity.

9.10 Theorem. *Let there exist a solution* x_* *of the problem* (P) *in the convex case, where the Slater condition is satisfied. We define the value function* V *on* $\mathbb{R}^m \times \mathbb{R}^n$ *as follows*:

$$V(\alpha,\beta) = \inf\left\{ f(x) : x \in S,\ g(x) \leqslant \alpha,\ h(x) = \beta \right\}.$$

Then V *is a convex function with values in* $(-\infty,+\infty]$. *The vector* $(1,\gamma,\lambda)$ *is a multiplier associated with* x_* *if and only if* $(\gamma,\lambda) \in -\partial V(0,0)$.

Proof. According to Theorems 9.4 and 9.8, there exists a normal multiplier $(1,\gamma,\lambda)$ associated to x_*. Let $\mathrm{P}(\alpha,\beta)$ denote the optimization problem that defines the value of $V(\alpha,\beta)$, and let x be any point admissible for $\mathrm{P}(\alpha,\beta)$:

$$x \in S,\ g(x) \leqslant \alpha,\ h(x) = \beta.$$

Then, using $\gamma \geqslant 0$ and $-g(x) \geqslant -\alpha$, we have

$$\begin{aligned} f(x) &= f(x) + \langle \gamma, g(x) \rangle + \langle \gamma, -g(x) \rangle + \langle \lambda, h(x) - \beta \rangle \\ &\geqslant f(x) + \langle \gamma, g(x) \rangle + \langle \gamma, -\alpha \rangle + \langle \lambda, h(x) - \beta \rangle \\ &\geqslant f(x_*) - \langle \gamma, \alpha \rangle - \langle \lambda, \beta \rangle = V(0,0) - \langle (\gamma,\lambda), (\alpha,\beta) \rangle, \end{aligned}$$

by the minimization condition of the multiplier $(1,\gamma,\lambda)$. Taking the infimum over x, we deduce

$$V(\alpha,\beta) \geqslant V(0,0) - \langle (\gamma,\lambda), (\alpha,\beta) \rangle,$$

which confirms $V > -\infty$, and that $-(\gamma,\lambda)$ belongs to $\partial V(0,0)$, the subdifferential of V at $(0,0)$. As for the convexity of V, it follows easily from its definition (or it can be deduced from Exer. 8.10).

As the reader well knows, it is not our custom to abandon a proof in midstream. On this occasion, however, we would ask the reader to kindly supply the converse; it happens to be the subject of the exercise that follows. □

9.11 Exercise. Under the hypotheses of Theorem 9.10, prove that an element (γ,λ) belonging to $-\partial V(0,0)$ determines a normal multiplier $(1,\gamma,\lambda)$. □

9.3 Convex duality

An important feature of convex optimization is the possibility of developing a theory in which one associates to the original, or *primal*, problem another optimization problem, the *dual*, which is linked to the primal through multipliers (of some type or other). This idea has important theoretical and even numerical consequences, in such areas as game theory, optimal transport, operations research, mechanics, and

economics. We take a brief look at the topic in this section, in order to establish its connection to the multiplier rule.

We continue to be interested in the problem (P) of the preceding section, which we think of as the primal. The **dual problem** (D) associated to (P) is defined as follows:

$$\text{Maximize } \varphi(\gamma,\lambda) \text{ subject to } (\gamma,\lambda) \in \mathbb{R}^m_+ \times \mathbb{R}^n \tag{D}$$

where the (concave) function $\varphi : \mathbb{R}^m \times \mathbb{R}^n \to [-\infty,+\infty)$ is defined by

$$\varphi(\gamma,\lambda) = \inf_{x \in S} \big\{ f + \langle \gamma, g \rangle + \langle \lambda, h \rangle \big\}(x).$$

The dual problem is of greatest interest when it can be rendered more explicit. Let's illustrate this now.

9.12 Example. We are given $c \in \mathbb{R}^n$, $b \in \mathbb{R}^m$, and a matrix M which is $m \times n$, and we consider the following instance of the problem (P):

$$\text{Minimize } \langle c, x \rangle \text{ subject to } x \in \mathbb{R}^n_+, \quad Mx \leqslant b.$$

(As usual, points in Euclidean space, in their dealings with matrices, are viewed as columns.) This is a problem in what is called *linear programming*. We are dealing, then, with the convex case of (P), in the absence of equality constraints. Let us make explicit the dual problem (D). We have

$$\begin{aligned}\varphi(\gamma) &= \inf \big\{ \langle c, x \rangle + \langle \gamma, Mx - b \rangle : x \in \mathbb{R}^n_+ \big\} \\ &= \inf \big\{ \langle c + M^*\gamma, x \rangle - \langle \gamma, b \rangle : x \in \mathbb{R}^n_+ \big\} = \begin{cases} -\langle \gamma, b \rangle & \text{if } c + M^*\gamma \geqslant 0 \\ -\infty & \text{otherwise.} \end{cases}\end{aligned}$$

It turns out then, that (D) can be expressed as follows:

$$\text{Maximize } \langle -b, \gamma \rangle \text{ subject to } \gamma \in \mathbb{R}^m_+, \quad -M^*\gamma \leqslant c.$$

Thus, the dual problem has essentially the same form as the primal; this fact is exploited to great effect in the subject. □

We now describe the link between the primal and the dual problem.

9.13 Theorem. (Lagrangian duality) *We consider the basic problem* (P) *in the convex case. We suppose that there is a solution of* (P) *which admits a normal multiplier. Then*

$$\min(\text{P}) = \max(\text{D}).$$

Furthermore, any solution x_* of (P) *admits a normal multiplier, and a vector $(1, \gamma_*, \lambda_*) \in \mathbb{R} \times \mathbb{R}^m_+ \times \mathbb{R}^n$ is a multiplier for x_* if and only if (γ_*, λ_*) solves the dual problem* (D).

Proof.

A. Let $(\gamma,\lambda) \in \mathbb{R}^m_+ \times \mathbb{R}^n$. Observe that

$$\varphi(\gamma,\lambda) \leqslant \{f+\langle \gamma, g\rangle + \langle \lambda, h\rangle\}(x_*) \leqslant f(x_*) = \min(\mathrm{P}).$$

It follows that sup (D) $\leqslant$ min (P). Now let $(1,\gamma_*,\lambda_*)$ be a multiplier for a solution x_* of (P). (By hypothesis, at least one such normal multiplier and solution exist.) The minimization condition asserts

$$\{f+\langle \gamma_*, g\rangle + \langle \lambda_*, h\rangle\}(x) \geqslant \{f+\langle \gamma_*, g\rangle + \langle \lambda_*, h\rangle\}(x_*) = f(x_*) \;\; \forall x \in S,$$

whence

$$\sup(\mathrm{D}) \geqslant \varphi(\gamma_*,\lambda_*) \geqslant f(x_*) = \min(\mathrm{P}) \geqslant \sup(\mathrm{D}).$$

We deduce that (γ_*,λ_*) solves the dual problem, and that min (P) = max (D).

B. Now let (γ_*,λ_*) be any solution of the dual problem, and let x_* be any solution of the primal problem. Then $\gamma_* \in \mathbb{R}^m_+$, and we have

$$\sup(\mathrm{D}) = \varphi(\gamma_*,\lambda_*) \leqslant \{f+\langle \gamma_*, g\rangle + \langle \lambda_*, h\rangle\}(x_*) \leqslant f(x_*) = \min(\mathrm{P}) = \sup(\mathrm{D}),$$

which implies $\langle \gamma_*, g(x_*)\rangle = 0$, the complementary slackness condition. We also have, for any $x \in S$,

$$\{f+\langle \gamma_*, g\rangle + \langle \lambda_*, h\rangle\}(x) \geqslant \varphi(\gamma_*,\lambda_*) = \sup(\mathrm{D}) = \min(\mathrm{P}) = f(x_*),$$

which yields the minimization condition for $(1,\gamma_*,\lambda_*)$, and confirms that this vector has all the properties of a multiplier for x_*. □

9.14 Exercise. Under the hypotheses of Theorem 9.13, let x_* be a solution of (P), and let $(1,\gamma_*,\lambda_*)$ be a (normal) multiplier associated to x_*. The *Lagrangian* L of the problem is defined to be the function

$$L(x,\gamma,\lambda) = \{f+\langle \gamma, g\rangle + \langle \lambda, h\rangle\}(x).$$

Prove that (x_*,γ_*,λ_*) is a *saddle point* of L, meaning that

$$L(x_*,\gamma,\lambda) \leqslant L(x_*,\gamma_*,\lambda_*) \leqslant L(x,\gamma_*,\lambda_*) \;\; \forall x \in S,\; \gamma \in \mathbb{R}^m_+,\; \lambda \in \mathbb{R}^n.$$ □

Remark. If x is admissible for (P), we know, of course, that min (P) $\leqslant f(x)$. Similarly, if (γ,λ) is admissible for (D), we obtain max $(D) \geqslant \varphi(\gamma,\lambda)$. But now, suppose that duality holds: min (P) = max (D). Then we deduce

$$\varphi(\gamma,\lambda) \leqslant \min(\mathrm{P}) \leqslant f(x).$$

The generating of *bilateral* bounds of this type is of evident interest in developing numerical methods, a task to which duality has been effectively applied. Under more subtle hypotheses than those of Theorem 9.13 (in linear programming, or in

the presence of infinite-dimensional equality constraints, for example), it can be a delicate matter to establish duality.

Another salient point, as evidenced by Theorem 9.10, is the possibility of finding the multipliers for the primal problem by solving the dual problem. We illustrate this procedure now.

9.15 Exercise. We are given n continuous, convex functions $f_i : [0,\infty) \to \mathbb{R}$, and a positive parameter q. We study the following simple *allocation* problem, of a type that frequently arises in economics and operations research:

$$\text{Minimize } f(x) = \sum_{i=1}^{n} f_i(x_i) \text{ subject to } x \in S := \mathbb{R}^n_+, \quad \sum_{i=1}^{n} x_i \leqslant q. \qquad \text{(P)}$$

Note that this is a convex case of (P), with no equality constraints.

The i-th cost component f_i depends only on x_i; the difficulty (especially when n is large) lies in determining what optimal $x_i \geqslant 0$ to allocate to f_i , while respecting the upper bound on the sum of the x_i.

(a) Prove that a solution x_* exists, verify the Slater condition, and deduce that x_* admits a normal multiplier $(1, \gamma_*)$.

(b) Prove that, for each index i, the value x_{*i} is a solution of the problem

$$\min \left\{ f_i(u) + \gamma_* u \;:\; u \in \mathbb{R}_+ \right\}.$$

It turns out then, that if we know γ_*, we may use it to calculate x_* one coordinate at a time, while completely ignoring the constraint $\sum_i x_i \leqslant q$. The problem is said to have been *decomposed*.[4]

How might we effectively calculate γ_*, however? For this purpose, we define the following function closely related to the conjugate of f_i:

$$\theta_i(\gamma) = \inf \left\{ \gamma u + f_i(u) : u \geqslant 0 \right\}, \;\; \gamma \in \mathbb{R}.$$

(c) Show that the dual problem (D) consists of maximizing over $\mathbb{R}_+$ the following function of a single variable:

$$\varphi(\gamma) = \sum_{i=1}^{n} \theta_i(\gamma) - \gamma q.$$

Why is this problem relatively simple? How do we know a maximum exists?

We suppose henceforth, in order to permit explicit calculation, that each f_i has the form

$$f_i(u) = p_i u + u \ln u \text{ for } u > 0, \text{ with } f_i(0) = 0.$$

[4] Thus, the computation could be envisaged on a *decentralized* basis, where each component, having been informed of the internal unit cost γ_*, can calculate its own allocation x_{*i} by maximizing its own profit. These individually motivated calculations would lead to *global* optimality: Adam Smith's invisible hand at work.

(d) Let $\psi : \mathbb{R}_+ \to \mathbb{R}$ be defined by

$$\psi(u) = u \ln u \text{ for } u > 0,\ \psi(0) = 0,$$

and let $c \in \mathbb{R}$. Prove that the function $u \mapsto \psi(u) - cu$ attains a minimum over $\mathbb{R}_+$ at $u = e^{c-1}$, the corresponding value of ψ being $-e^{c-1}$.

e) Deduce from this the evident solution to problem (P) when q is no less than

$$\sigma := \sum_{i=1}^{n} e^{-p_i - 1}.$$

Prove that when $q < \sigma$, the solution γ_* of the dual problem is $\ln(\sigma/q)$. Use this to show that the optimal allocation is given by $x_{*i} = e^{-p_i-1} q/\sigma$.

(f) Prove that the value $V(q)$ of the problem (P) is given by

$$V(q) = \begin{cases} -\sigma & \text{if } q \geqslant \sigma \\ q\big(\ln q - 1 - \ln \sigma\big) & \text{if } q \leqslant \sigma. \end{cases}$$

Show that $V'(q) = -\gamma_*$ (the expected sensitivity relation). □

Chapter 10

Generalized gradients

The goal of this chapter is to develop a generalized calculus on Banach spaces, one that reduces to differential calculus for smooth functions, and to subdifferential calculus for convex functions. A natural question arises: what class of functions to consider? This leads us to ponder what smooth functions and convex functions have in common. One answer to this is: the local Lipschitz property. If $f : X \to \mathbb{R}$ is a continuously differentiable function on a Banach space X, then, as we know, f is locally Lipschitz. If, instead, f is convex, then once again it has that property, as we have seen (Theorem 5.17). We choose, therefore, to work in the class of *locally Lipschitz functions*.

This class has other features that recommend it as an environment for the task. It is closed under familiar operations such as sum, product, and composition. But it is also closed under less classical ones, such as taking lower or upper envelopes. Finally, the class of locally Lipschitz functions includes certain nonsmooth, nonconvex functions that are important in a variety of applications. A notable example is provided by distance functions.

Figure 10.1 below shows the graph of a distance function d_S, where S is the union of two segments in $\mathbb{R}$. This simple example illustrates the fact that distance functions are generally neither smooth nor convex.

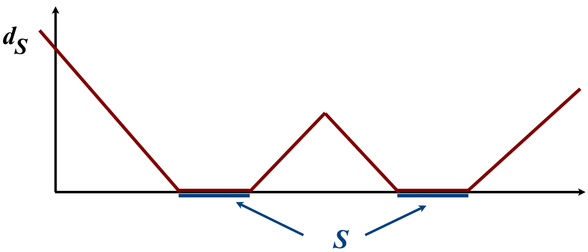


Fig. 10.1 The graph of a distance function d_S

The next example of nonsmoothness arises in optimization.

F. Clarke, *Functional Analysis, Calculus of Variations and Optimal Control*, Graduate Texts in Mathematics 264, DOI 10.1007/978-1-4471-4820-3_10,

10.1 Example. (Robust optimization) Many situations in applied optimization involve uncertainties that must be taken into account. In designing a product, engineers may have at their disposal a design parameter x, and the goal may be to minimize a cost $f(x)$. However, a perturbation q may intervene in the manufacture. It is known, we suppose, that q lies in a certain compact set Q, but nothing else. We denote the resulting inaccuracy by $e(x,q)$, and we suppose that this must be no greater than a specified acceptable level E. We assume that e is continuously differentiable.

Since q is unknown, the design parameter x must be chosen so that $g(x) \leqslant 0$, where g is defined by

$$g(x) \,=\, \max_{q \in Q} e(x,q) - E\,.$$

The function g is unlikely to be differentiable, even though e is smooth, and will fail in general to be convex. It is easy to show, however, that g is locally Lipschitz. Thus the minimization of f is subject to an inequality constraint specified by a nonsmooth, nonconvex function.

This type of problem has been one of the motivations to consider optimization with nonsmooth and nonconvex data, among others that arise in such areas as optimal design, eigenvalue placement, and elasticity. Note that the classical multiplier rule (Theorem 9.1) cannot be applied here unless Q is a finite set. We shall see later (in Exer. 10.26) how to obtain necessary conditions for this problem via nonsmooth calculus. □

In this chapter, we develop the calculus of the *generalized gradient* of a locally Lipschitz function f, denoted $\partial_C f(x)$. This leads to a *unified* treatment of smooth and convex calculus. It also gives rise to an associated geometric theory of tangents and normals to arbitrary closed sets.

10.1 Definition and basic properties

Throughout this chapter, X denotes a Banach space. Let $f\colon X \to \mathbb{R}$ be *Lipschitz of rank K* near a given point $x \in X$; that is, for some $\varepsilon > 0$, we have

$$|\, f(y) - f(z)\,| \,\leqslant\, K \|\, y - z\,\| \;\; \forall\, y, z \in B(x,\varepsilon)\,.$$

The *generalized directional derivative* of f at x in the direction v, denoted $f^{\circ}(x\,;v)$, is defined as follows:

$$f^{\circ}(x\,;v) \,=\, \limsup_{\substack{y \to x \\ t \downarrow 0}} \frac{f(y+tv) - f(y)}{t}\,,$$

where y lives in X and t is a positive scalar. Note that this definition does not presuppose the existence of any limit (since it involves an upper limit only), that it involves only the behavior of f arbitrarily near x, and that it differs from the traditional definition of the directional derivative in that the base point y of the difference quotient varies. The utility of f° stems from the basic properties recorded below. We recall that a function g is said to be positively homogeneous if $g(\lambda v) = \lambda g(v)$ for $\lambda \geqslant 0$, and subadditive if $g(v+w) \leqslant g(v)+g(w)$ for all v,w.

10.2 Proposition. *Let f be Lipschitz of rank K near x. Then:*

(a) *The function $v \mapsto f^{\circ}(x;v)$ is finite, positively homogeneous, and subadditive on X, and satisfies $|f^{\circ}(x;v)| \leqslant K\|v\|,\ v \in X$;*

(b) *For every $v \in X$, the function $(u,w) \mapsto f^{\circ}(u;w)$ is upper semicontinuous at (x,v); the function $w \mapsto f^{\circ}(x;w)$ is Lipschitz of rank K on X;*

(c) *We have $f^{\circ}(x;-v) = (-f)^{\circ}(x;v),\ v \in X$.*

Proof. In view of the Lipschitz condition, the absolute value of the difference quotient in the definition of $f^{\circ}(x;v)$ is bounded by $K\|v\|$ when y is sufficiently near x and t sufficiently near 0. It follows that $|f^{\circ}(x;v)|$ admits the same upper bound. The fact that $f^{\circ}(x;\lambda v) = \lambda f^{\circ}(x;v)$ for any $\lambda \geqslant 0$ is immediate, so let us turn now to the subadditivity. With all the upper limits below understood to be taken as $y \to x$ and $t \downarrow 0$, we calculate:

$$\begin{aligned} f^{\circ}(x;v+w) &= \limsup \frac{f(y+tv+tw)-f(y)}{t} \\ &\leqslant \limsup \frac{f(y+tv+tw)-f(y+tw)}{t} + \limsup \frac{f(y+tw)-f(y)}{t} \end{aligned}$$

(since the upper limit of a sum is bounded above by the sum of the upper limits). The first upper limit in this last expression is $f^{\circ}(x;v)$, since the term $y+tw$ represents in essence just a dummy variable converging to x. We conclude

$$f^{\circ}(x;v+w) \leqslant f^{\circ}(x;v)+f^{\circ}(x;w).$$

which establishes (a).

Now let x_i and v_i be arbitrary sequences converging to x and v, respectively. For each i, by definition of the upper limit, there exist y_i in X and $t_i > 0$ such that $\|y_i - x_i\| + t_i < 1/i$ and

$$\begin{aligned} f^{\circ}(x_i;v_i) - \frac{1}{i} &\leqslant \frac{f(y_i+t_iv_i)-f(y_i)}{t_i} \\ &= \frac{f(y_i+t_iv)-f(y_i)}{t_i} + \frac{f(y_i+t_iv_i)-f(y_i+t_iv)}{t_i}. \end{aligned}$$

Note that the last term is bounded in magnitude by $K\|v_i - v\|$ (in view of the Lipschitz condition). Upon taking upper limits (as $i \to \infty$), we derive

$$\limsup_{i\to\infty} f^{\circ}(x_i;v_i) \leqslant f^{\circ}(x;v),$$

which establishes the upper semicontinuity claimed in (b). Now let any v and w in X be given. By the Lipschitz property, we have

$$f(y+tv) - f(y) \leqslant f(y+tw) - f(y) + K\|v-w\|t$$

for all y near x and positive t near 0. Dividing by t and taking upper limits as $y \to x$, $t \downarrow 0$ leads to

$$f^{\circ}(x;v) \leqslant f^{\circ}(x;w) + K\|v-w\|.$$

Since this also holds with v and w switched, the remaining assertion of (b) follows. To prove (c), we calculate:

$$\begin{aligned} f^{\circ}(x;-v) &= \limsup_{\substack{y\to x\\ t\downarrow 0}} \frac{f(y-tv)-f(y)}{t} \\ &= \limsup_{\substack{u\to x\\ t\downarrow 0}} \frac{(-f)(u+tv)-(-f)(u)}{t}, \text{ where } u := y-tv \\ &= (-f)^{\circ}(x;v), \end{aligned}$$

which confirms the stated formula. □

A function such as $v \mapsto f^{\circ}(x;v)$ which is positively homogeneous and subadditive on X, and bounded on the unit ball, is the support function of a uniquely determined weak* compact convex set in X^*, as we have seen in Theorem 4.25.

10.3 Definition. *The* **generalized gradient** *of the function f at x, denoted $\partial_C f(x)$, is the unique nonempty weak* compact convex subset*[1] *of X^* whose support function is $f^{\circ}(x;\cdot)$. We have therefore*

$$\zeta \in \partial_C f(x) \iff f^{\circ}(x;v) \geqslant \langle \zeta, v\rangle \ \forall v \in X,$$

$$f^{\circ}(x;v) = \max\left\{\langle \zeta, v\rangle : \zeta \in \partial_C f(x)\right\} \ \forall v \in X.$$

10.4 Exercise. Let f be Lipschitz near x, and Gâteaux differentiable at x. Prove that $f'_G(x)$ belongs to $\partial_C f(x)$. □

10.5 Proposition. *Let f be Lipschitz of rank K near x. Then $\partial_C f(x) \subset B_*(0,K)$.*

Proof. For any $\zeta \in \partial_C f(x)$, we have $\langle \zeta, v\rangle \leqslant K\|v\| \ \forall v \in X$, by Prop. 10.2. □

---

[1] The compact convex set $\partial_C f$ is often pronounced "dee cee eff".

10.6 Example. Let $f(x) = \|x\|$, a function which is Lipschitz of rank 1. It follows from Prop. 10.2 that we have $f^{\circ}(0;v) \leqslant \|v\| \ \forall v \in X$. But $f'(0;v)$, the usual directional derivative, is given by $\|v\|$, and of course we have $f^{\circ}(0;v) \geqslant f'(0;v)$. These facts yield:

$$f^{\circ}(0;v) = \|v\| \ \forall v \in X.$$

Now we may proceed to calculate $\partial_C f(0)$ from Def. 10.3, according to which ζ belongs to $\partial_C f(0)$ if and only if

$$f^{\circ}(0;v) = \|v\| \geqslant \langle \zeta, v \rangle \ \forall v \in X.$$

We discover $\partial_C f(0) = B_*(0,1)$. We remark that, just as derivatives are rarely calculated directly from difference quotients in practice, so too will the calculation of generalized gradients generally be done with the help of calculus rules. □

10.7 Exercise. Let $f : X \to \mathbb{R}$ be locally Lipschitz.

(a) Prove Fermat's rule for ∂_C: If f has a local minimum or maximum at x, then $0 \in \partial_C f(x)$.

(b) A unit vector $v \in X$ is said to be a *descent direction* for f at x if

$$\limsup_{t \downarrow 0} \frac{f(x+tv) - f(x)}{t} < 0.$$

Prove that if $0 \notin \partial_C f(x)$, then a descent direction exists.

(c) Let X be a Hilbert space, and suppose that $0 \notin \partial_C f(x)$. Let ζ be the element of minimal norm in $\partial_C f(x)$. Prove that $v := -\zeta / |\zeta|$ is a descent direction satisfying

$$\limsup_{t \downarrow 0} \frac{f(x+tv) - f(x)}{t} \leqslant -d\big(0, \partial_C f(x)\big) < 0.$$

Such descent directions play an important role in numerical algorithms for nonsmooth optimization. □

Smooth and convex functions. A function $f : X \to \mathbb{R}$ which is continuously differentiable near a point x is locally Lipschitz near x, by the mean value theorem. A function $f : X \to \mathbb{R}_\infty$ which is convex and lsc is locally Lipschitz in $\operatorname{int} \operatorname{dom} f$ (Theorem 5.17). In both these cases, as we now see, $\partial_C f$ reduces to the familiar concept: the derivative or the subdifferential.

10.8 Theorem. *If f is continuously differentiable near x, then $\partial_C f(x) = \{f'(x)\}$. If f is convex and lsc, and if $x \in \operatorname{int} \operatorname{dom} f$, then $\partial_C f(x) = \partial f(x)$.*

Proof. In both cases addressed by the theorem, one is asserting an equality between two convex, weak* compact sets. We may establish it by showing that their support

functions coincide, in view of Cor. 3.13. The support function of the set appearing on the left (in either assertion) is $f^\circ(x;v)$, by definition of $\partial_C f(x)$.

In the smooth case, the support function of the set on the right (evaluated at v) is $\langle f'(x), v\rangle = f'(x;v)$. Thus, we must show that $f^\circ(x;v) = f'(x;v)$. Let $y_i \to x$ and $t_i \downarrow 0$ be sequences realizing $f^\circ(x;v)$, in the sense that

$$f^\circ(x;v) = \lim_{i\to\infty} \frac{f(y_i + t_i v) - f(y_i)}{t_i}.$$

Then we have

$$f^\circ(x;v) = \lim_{i\to\infty} \frac{f(y_i + t_i v) - f(y_i)}{t_i} = \lim_{i\to\infty} \langle f'(z_i), v\rangle$$

(for $z_i \in [y_i, y_i + t_i v]$, by the mean value theorem)

$$= \langle f'(x), v\rangle = f'(x;v),$$

since f' is continuous. The first part of the theorem follows.

We turn now to the convex case. We know that $f'(x;\cdot)$ is the support function of $\partial f(x)$, by Cor. 4.26. Now $f'(x;v) \leqslant f^\circ(x;v)$ by the way these are defined; it suffices therefore to prove the opposite inequality. Fix $\delta > 0$. Then

$$\begin{aligned} f^\circ(x;v) &= \lim_{\varepsilon\downarrow 0} \sup_{\|y-x\|\leqslant \delta\varepsilon} \sup_{0<t<\varepsilon} \frac{f(y+tv) - f(y)}{t} \\ &= \lim_{\varepsilon\downarrow 0} \sup_{\|y-x\|\leqslant \delta\varepsilon} \frac{f(y+\varepsilon v) - f(y)}{\varepsilon} \end{aligned}$$

(since $t \mapsto [f(y+tv) - f(y)]/t$ is increasing; see Prop. 2.22)

$$\leqslant \lim_{\varepsilon\downarrow 0} \frac{f(x+\varepsilon v) - f(x)}{\varepsilon} + 2\delta K$$

(where K is a Lipschitz constant for f in a neighborhood of x)

$$= f'(x;v) + 2\delta K.$$

Since $\delta > 0$ is arbitrary, this completes the proof. □

10.9 Exercise. Let f be Lipschitz near x, and suppose that $\partial_C f(x)$ is a singleton $\{\zeta\}$. Prove that f is Gâteaux differentiable at x, and that $f'_G(x) = \zeta$. □

We remark that $\partial_C f(x)$ can fail to be a singleton when f is differentiable at x (see Exer. 13.10). This reflects the fact that the generalized gradient extends the notion of *continuous* differentiability, as evidenced by the following.

10.10 Proposition. *Let f be Lipschitz of rank K near x, and let x_i and ζ_i be sequences in X and X^* such that*

$$x_i \to x \quad \text{and} \quad \zeta_i \in \partial_C f(x_i) \ \forall\, i.$$

If ζ is a weak* cluster point of the sequence ζ_i, (in particular, if ζ_i converges to ζ in X^*), then we have $\zeta \in \partial_C f(x)$.

Proof. Fix $v \in X$. For each i, we have $f^\circ(x_i;v) \geqslant \langle \zeta_i, v\rangle$. The sequence $\langle \zeta_i, v\rangle$ is bounded in $\mathbb{R}$, and contains terms that are arbitrarily near $\langle \zeta, v\rangle$. Let us extract a subsequence of ζ_i (without relabeling) such that $\langle \zeta_i, v\rangle \to \langle \zeta, v\rangle$. Then passing to the limit in the preceding inequality, and using the fact that f° is upper semicontinuous in x (Prop. 10.2), we deduce

$$f^\circ(x;v) \geqslant \langle \zeta, v\rangle.$$

Since v is arbitrary, it follows that $\zeta \in \partial_C f(x)$. ☐

10.2 Calculus of generalized gradients

The reader, like many mathematicians, may have first learned to love mathematics by doing calculus. We now have an opportunity to do calculus all over again.

10.11 Proposition. *For any scalar λ, we have $\partial_C(\lambda f)(x) = \lambda\, \partial_C f(x)$.*

Proof. Note that the function λf is Lipschitz near x. When λ is nonnegative, then $(\lambda f)^\circ(x;\cdot) = \lambda f^\circ(x;\cdot)$, and the result follows, since these are the support functions of the two convex, weak* compact sets involved. To complete the proof, it suffices to consider now the case $\lambda = -1$. An element ζ of X^* belongs to $\partial_C(-f)(x)$ if and only if $(-f)^\circ(x;v) \geqslant \langle \zeta, v\rangle$ for all v. By Proposition 10.2 (c), this is equivalent to: $f^\circ(x;-v) \geqslant \langle \zeta, v\rangle$ for all v, which is equivalent to $-\zeta$ belonging to $\partial_C f(x)$, by definition of $\partial_C f(x)$. ☐

We now pause for a moment in order to introduce a certain functional property that will be useful in our development of generalized gradient calculus.

10.12 Definition. *We say that f is* **regular** *at x provided f is Lipschitz near x and admits directional derivatives $f'(x;v)$ satisfying $f^\circ(x;v) = f'(x;v) \ \forall\, v \in X$.*

Let us point out that, as showed in the proof of Theorem 10.8, a continuously differentiable function is regular at any point, as is a lower semicontinuous convex function (at points in the interior of its effective domain).

On the other hand, let f be a continuous *concave* function that has a corner at x; that is, such that for some v, we have $f'(x;v) \neq -f'(x;-v)$. Then f fails to be regular at x. To see this, bear in mind that $-f$ is convex and therefore regular, whence

$$f'(x;v) \neq -f'(x;-v) = (-f)'(x;-v) = (-f)^{\circ}(x;-v) = f^{\circ}(x;v).$$

Thus $f^{\circ}(x;v) \neq f'(x;v)$, so that f is not regular at x. Roughly speaking, we can think of regular functions as those that, at each point, are either smooth, or else have a corner of convex type.

Certain general calculus rules become "more exact" when the underlying functions are regular, as illustrated by the following.

10.13 Theorem. (Sum rule) *Let f and g be Lipschitz near x. Then*

$$\partial_C(f+g)(x) \subset \partial_C f(x) + \partial_C g(x).$$

Equality holds if f and g are regular at x.

Proof. Both sides of the inclusion are weak* closed convex subsets (see Exer. 8.8), so (for the first assertion of the theorem) it suffices, by Cor. 3.13, to verify the following inequality between support functions: $H_1 \leqslant H_2$, where H_1 is the support function of the set on the left, and H_2 of the set on the right. Now $H_1(v)$ is $(f+g)^{\circ}(x;v)$, and $H_2(v)$ is $f^{\circ}(x;v)+g^{\circ}(x;v)$. The following lemma therefore completes the proof of the first assertion.

Lemma. $(f+g)^{\circ}(x;v) \leqslant f^{\circ}(x;v)+g^{\circ}(x;v) \;\; \forall\, v \in X.$

This follows directly from

$$\limsup_{\substack{y \to x \\ t \downarrow 0}} \frac{(f+g)(y+tv)-(f+g)(y)}{t} \leqslant \limsup_{\substack{y \to x \\ t \downarrow 0}} \frac{f(y+tv)-f(y)}{t} + \limsup_{\substack{y \to x \\ t \downarrow 0}} \frac{g(y+tv)-g(y)}{t}.$$

When f and g are regular at x, the inequality of the lemma becomes an equality, since

$$f^{\circ}(x;v)+g^{\circ}(x;v) = f'(x;v)+g'(x;v) = (f+g)'(x;v) \leqslant (f+g)^{\circ}(x;v).$$

Then equality holds between the sets as well. ☐

We obtain a simple example of strict containment in the sum rule as follows: take $f(x) = \|x\|$ and $g(x) = -\|x\|$. Then

$$\{0\} = \partial_C(f+g)(0) \subsetneqq \partial_C f(0) + \partial_C g(0) = 2B_*(0,1).$$

The following exercise implies, for example, that the sum of a continuously differentiable function and a convex function is regular.

10.14 Exercise. Show that a positive linear combinations of functions regular at x is regular at x. □

The extension of the sum rule to finite linear combinations is immediate, in view of Proposition 10.11 and Exer. 10.14:

10.15 Proposition. *Let f_i be Lipschitz near x, and let λ_i be scalars $(i = 1, 2, \ldots, n)$. Then $f := \sum_{i=1}^n \lambda_i f_i$ is Lipschitz near x, and we have*

$$\partial_C\Big(\sum_{i=1}^n \lambda_i f_i\Big)(x) \subset \sum_{i=1}^n \lambda_i \partial_C f_i(x).$$

Equality holds if, for each i, f_i is regular at x and $\lambda_i \geqslant 0$.

10.16 Exercise. Let f be Lipschitz near x, and let g be continuously differentiable near x. Prove that

$$\partial_C(f+g)(x) = \partial_C f(x) + \{g'(x)\}.$$

□

Just as Lagrange's mean value theorem is a basic tool in classical calculus, the following result (due to Lebourg) is of frequent use.

10.17 Theorem. (Mean value theorem) *Let x and y belong to X, and suppose that f is Lipschitz on a neighborhood of the line segment $[x,y]$. Then there exists a point z in (x,y) such that*

$$f(y) - f(x) \in \langle \partial_C f(z), y-x \rangle.$$

Proof. We will need the following special case of the chain rule for the proof. We denote by x_t the point $x + t(y-x)$.

Lemma. *The function $g\colon [0,1] \to \mathbb{R}$ defined by $g(t) = f(x_t)$ is Lipschitz on $(0,1)$, and we have $\partial_C g(t) \subset \big\langle \partial_C f(x_t), y-x \big\rangle$.*

Proof. That g is Lipschitz is clear. The two closed convex sets appearing in the inclusion are in fact intervals in $\mathbb{R}$, so it suffices to prove that for $v = \pm 1$, we have

$$\max\big\{\partial_C g(t) v\big\} \leqslant \max\big\{\langle \partial_C f(x_t), y-x\rangle v\big\}.$$

Now the left-hand side of this inequality is just $g^\circ(t;v)$, by the definition of $\partial_C g(t)$. Writing out in turn the definition of $g^\circ(t;v)$, we calculate

$$\limsup_{\substack{s\to t\\ \lambda\downarrow 0}} \frac{g(s+\lambda v)-g(s)}{\lambda} = \limsup_{\substack{s\to t\\ \lambda\downarrow 0}} \frac{f(x+[s+\lambda v](y-x))-f(x+s(y-x))}{\lambda}$$
$$\leqslant \limsup_{\substack{z\to x_t\\ \lambda\downarrow 0}} \frac{f(z+\lambda v(y-x))-f(z)}{\lambda}$$
$$= f^{\circ}\big(x_t\,;v(y-x)\big) = \max\big\langle\, \partial_C f(x_t),\, v(y-x)\,\big\rangle,$$

which completes the proof of the lemma. □

Now for the proof of the theorem. Consider the (continuous) function θ on $[0,1]$ defined by

$$\theta(t) = f(x_t)+t\,[\,f(x)-f(y)]\,.$$

Note that $\theta(0)=\theta(1)=f(x)$, so that there is a point t in $(0,1)$ at which θ attains a local minimum or maximum (by continuity). By Exer. 10.7, we have $0\in\partial_C\theta(t)$. We may calculate $\partial_C\theta(t)$ by appealing to Propositions 10.11 and 10.15, and the lemma. We deduce

$$0 \in f(x)-f(y)+\langle\,\partial_C f(x_t),\, y-x\rangle,$$

which is the assertion of the theorem (take $z=x_t$). □

10.18 Exercise. Let K be a nonempty cone in X, and let $f : X\to\mathbb{R}$ be locally Lipschitz. We say that f is decreasing relative to K provided that, for any $x\in X$, $y\in x+K \implies f(y)\leqslant f(x)$. Prove that f has this property if and only if, for every $x\in X$, we have

$$H_K(\zeta) := \max_{v\,\in\,K}\,\langle\,\zeta, v\rangle \leqslant 0 \;\;\forall\,\zeta\in\partial_C f(x)\,.$$
□

10.19 Theorem. (Chain rule 1) *Let Y be a Banach space, and let $F\colon X\to Y$ be continuously differentiable near x. Let $g\colon Y\to\mathbb{R}$ be Lipschitz near $F(x)$. Then the function $f := g\circ F$ is Lipschitz near x, and we have*

$$\partial_C f(x) \subset F'(x)^*\partial_C\, g\big(F(x)\big),$$

where * denotes the adjoint. If $F'(x)\colon X\to Y$ is onto, then equality holds.

Proof. The fact that f is Lipschitz near x is straightforward. In terms of support functions, we must prove that given any v, there is some element ζ of $\partial_C\, g(F(x))$ such that

$$f^{\circ}(x;v) \leqslant \langle\, v, F'(x)^*\zeta\,\rangle = \langle\,\zeta, F'(x)v\,\rangle.$$

Let the sequences $y_i\to x$ and $t_i\downarrow 0$ realize $f^{\circ}(x;v)$; that is

$$\lim_{i\to\infty} \frac{f(y_i+t_i v)-f(y_i)}{t_i} = f^{\circ}(x;v).$$

We have, for all i sufficiently large,

$$\frac{f(y_i+t_i v)-f(y_i)}{t_i} = \Big\langle \zeta_i\,,\, \frac{F(y_i+t_i v)-F(y_i)}{t_i} \Big\rangle$$

for some $\zeta_i \in \partial_C g(z_i)$, where z_i lies in the segment $(F(y_i), F(y_i+t_i v))$, by the mean value theorem 10.17. Since the sequence ζ_i is bounded (by a suitable Lipschitz constant for g), it admits a cluster point ζ in the weak* topology by Alaoglu's theorem (see Cor. 3.15).

We may extract a subsequence of ζ_i (without relabeling) for which $\langle \zeta_i, F'(x)v \rangle$ converges to $\langle \zeta, F'(x)v \rangle$. The required inequality now follows, using the fact that $[F(y_i+t_i v)-F(y_i)]/t_i$ converges (strongly) to $F'(x)v$.

Now suppose that $F'(x)$ is onto. Then F maps every neighborhood of x to a neighborhood of $F(x)$ (Cor. 5.33). This fact justifies the equality

$$\limsup_{\substack{y \to F(x)\\ t \downarrow 0}} \frac{g\big(y+tF'(x)v\big)-g(y)}{t} = \limsup_{\substack{u \to x\\ t \downarrow 0}} \frac{g\big(F(u)+tF'(x)v\big)-g\big(F(u)\big)}{t}.$$

Since $[F(u+tv)-F(u)-tF'(x)v]/t$ goes to zero as $u \to x$ and $t \downarrow 0$, and since g is Lipschitz locally, this leads to

$$\begin{aligned} g^\circ\big(F(x); F'(x)v\big) &= \limsup_{\substack{y \to F(x)\\ t \downarrow 0}} \frac{g\big(y+tF'(x)v\big)-g(y)}{t} \\ &= \limsup_{\substack{u \to x\\ t \downarrow 0}} \frac{g\big(F(u+tv)\big)-g\big(F(u)\big)}{t} = f^\circ(x;v). \end{aligned}$$

Since v is arbitrary, this implies equality between the two sets figuring in the statement of the theorem, as asserted. □

10.20 Theorem. (Chain rule 2) *Let $F: X \to \mathbb{R}^n$ be Lipschitz near x, and let the function $g: \mathbb{R}^n \to \mathbb{R}$ be continuously differentiable near $F(x)$. Then $f := g \circ F$ is Lipschitz near x, and we have*

$$\partial_C f(x) = \partial_C \langle g'\big(F(x)\big), F(\cdot)\rangle(x).$$

To clarify this equality, let us write γ for $g'(F(x))$. The right side above is $\partial_C h(x)$, where the real-valued function h is defined by $z \mapsto h(z) = \langle \gamma, F(z) \rangle$.

Proof. The fact that f is Lipschitz near x follows easily. We claim that it suffices to prove, for any $\varepsilon > 0$, the existence of a neighborhood V_ε of x in which we have

$$|f(y)-h(y)-(f(z)-h(z))| \leqslant \varepsilon|y-z| \;\; \forall y,z \in V_\varepsilon.$$

For then $\partial_C(f-h)(x)$ is contained in $B_*(0,\varepsilon)$ by Prop. 10.5, whence

$$\partial_C f(x) \subset \partial_C(f-h)(x) + \partial_C h(x) \subset \partial_C h(x) + B_*(0,\varepsilon),$$

by the sum rule 10.13. Switching f and h leads to $\partial_C h(x) \subset \partial_C f(x) + B_*(0,\varepsilon)$, and the required equality follows. To prove the claim, we observe that, for all y and z near x, we may write

$$\begin{aligned} f(y)-h(y)-f(z)+h(z) &= g(F(y)) - g(F(z)) - \langle \gamma, F(y)-F(z) \rangle \\ &= \langle g'(u)-\gamma, F(y)-F(z) \rangle \\ &\leqslant |g'(u)-\gamma|\,|F(y)-F(z)| \leqslant K_F\,|g'(u)-\gamma|\,\|y-z\|, \end{aligned}$$

where u lies between $F(y)$ and $F(z)$, and where K_F is a Lipschitz constant for F. It suffices to observe that by restricting y,z to a sufficiently small neighborhood of x, we can arrange to have $K_F\,|g'(u)-\gamma| < \varepsilon$. □

10.21 Exercise. Let f and g be Lipschitz near x. Then the product fg is Lipschitz near x. Use Theorem 10.20 to prove that

$$\partial_C\big(fg\big)(x) = \partial_C\big(f(x)\,g(\cdot) + g(x)\,f(\cdot)\big)(x) \subset f(x)\,\partial_C\,g(x) + g(x)\,\partial_C f(x).$$

Show that equality holds if f and g are regular at x and $f(x)\,g(x) \geqslant 0$. □

Upper envelopes. The taking of upper (or lower) envelopes, even of smooth functions, destroys smoothness; this is why there is no classical formula for the derivative of an envelope. However, taking envelopes preserves the Lipschitz property. We turn now to an early result in nonsmooth analysis that characterizes the directional derivatives of certain "max functions" defined this way.

Let Q be a compact metric space. Given an open subset V of X and a continuous function $g : V \times Q \to \mathbb{R}$, we define f on V as follows:

$$f(x) = \max_{q \in Q} g(x,q), \quad Q(x) = \big\{ q \in Q : f(x) = g(x,q) \big\}.$$

We suppose that $g'_x(x,q)$ (the derivative with respect to x) exists for every (x,q) in $V \times Q$, and that the mapping $(x,q) \mapsto g'_x(x,q)$ from $V \times Q$ to X^* (equipped with the dual norm) is continuous.

10.22 Theorem. (Danskin's formula) *f is regular at any point $x \in V$, and we have*

$$\partial_C f(x) = \overline{\text{co}}\,\big\{ g'_x(x,q) : q \in Q(x) \big\},$$

where the closed convex hull is taken in the weak sense. The directional derivatives of f are given by the formula*

$$f'(x;v) = \max_{q \in Q(x)} \langle g'_x(x,q), v \rangle, \ v \in X.$$

If $Q(x)$ is a singleton $\{q\}$, then $f'(x)$ exists and equals $g'_x(x,q)$.

Proof. We leave as an exercise the fact that f is locally Lipschitz in V (one uses the continuity of the derivative and Exer. 2.32).

Lemma 1.

$$f^{\circ}(x;v) \leqslant \max_{q \in Q(x)} \langle g_x'(x,q), v \rangle$$

We may suppose $v \neq 0$. Let x_i and t_i be sequences realizing $f^{\circ}(x;v)$, and let q_i belong to $Q(x_i + t_i v)$. Invoking the compactness of Q, we may suppose that $q_i \to q$; it follows that $q \in Q(x)$ (we ask the reader to show why). Then we may write, for some $z_i \in (x_i, x_i + t_i v)$,

$$\frac{f(x_i + t_i v) - f(x_i)}{t_i} \leqslant \frac{g(x_i + t_i v, q_i) - g(x_i, q_i)}{t_i} = \langle g_x'(z_i, q_i), v \rangle.$$

Passing to the limit, we obtain the inequality of Lemma 1.

Lemma 2.

$$\max_{q \in Q(x)} \langle g_x'(x,y), v \rangle \leqslant \liminf_{t \downarrow 0} \frac{f(x+tv) - f(x)}{t}.$$

To see this, let $q \in Q(x)$. Then

$$\frac{f(x+tv) - f(x)}{t} \geqslant \frac{g(x+tv, q) - g(x,q)}{t} = \langle g_x'(z,q), v \rangle$$

for some $z \in (x, x+tv)$. Taking lower limits leads to the stated inequality.

We turn now to the proof of the theorem. Since we always have

$$\limsup_{t \downarrow 0} \,(f(x+tv) - f(x))/t \leqslant f^{\circ}(x;v),$$

the lemmas immediately imply the stated formula, and that f is regular at x.

Suppose now that $Q(x)$ is a singleton $\{q\}$, and let x_i be a sequence converging to x ($x \neq x_i$). For every i sufficiently large, there exist $z_i \in (x, x_i)$ and $\zeta_i \in \partial_C f(z_i)$ such that $f(x_i) - f(x) = \langle \zeta_i, x_i - x \rangle$. Then

$$f(x_i) - f(x) \leqslant f^{\circ}(z_i; x_i - x) = \langle g_x'(z_i, q_i), x_i - x \rangle$$

for some $q_i \in Q(z_i)$ in view of the formula proved above, applied at z_i. A routine argument show that $q_i \to q$. We deduce

$$f(x_i) - f(x) - \langle g_x'(x,q), x_i - x \rangle \leqslant \langle g_x'(z_i, q_i) - g_x'(x,q), x_i - x \rangle,$$

whence

$$\limsup_{i \to \infty} \{ f(x_i) - f(x) - \langle g_x'(x,q), x_i - x \rangle \} / |x_i - x| \leqslant 0.$$

A symmetric argument using $-f^{\circ}(z_i\,;x-x_i)$ shows that the corresponding lim inf is nonnegative; it follows that $f'(x)$ exists and equals $g'_x(x,q)$. □

10.23 Corollary. (A special case) *Let $f_i : X \to \mathbb{R}$ $(i = 1,2,\dots,n)$ be continuously differentiable functions on an open set V. Define $f(x) = \max_{1 \leqslant i \leqslant n} f_i(x)$. Then f is regular at any point $x \in V$, and we have*

$$\partial_C f(x) = \operatorname{co}\big\{ f'_i(x) : i \in I(x) \big\}, \quad f'(x;v) = \max_{i \in I(x)} \langle f'_i(x), v \rangle \ \forall\, v,$$

where $I(x)$ denotes the set of indices i such that $f_i(x) = f(x)$.

This result gives a meaning (perhaps) to the phrase "the derivative of the max is the max of the derivatives."

10.24 Exercise. Prove the corollary. □

10.25 Example. Let us use the above to complete the stability analysis of systems of inequalities begun in Example 5.39. The missing step was to verify Hypothesis 5.30 for the locally Lipschitz function

$$\varphi(x,y) = \max\big\{0,\, g^1(x) - y^1,\, g^2(x) - y^2, \dots, g^m(x) - y^m\big\},$$

for some $\delta > 0$, and for (x,y) in a neighborhood of $(\bar{x},0)$. Theorem 10.22 assures us that φ has directional derivatives. We also know, as a consequence of the positive linear independence condition, that there exists a unit vector $\bar{v}$ such that

$$\langle Dg^i(\bar{x}), \bar{v} \rangle < 0 \ \ \forall\, i \in I(\bar{x}),$$

in view of Exer. 2.40.

Now suppose that Hypothesis 5.30 does not hold, and let us derive a contradiction. Since the hypothesis fails, there is a sequence $(x_j, y_j) \to (\bar{x},0)$ along which we have $\varphi(x_j, y_j) > 0$, and such that

$$\inf_{\|v\| \leqslant 1} \varphi'_x(x_j, y_j\,; v) > -1/j.$$

Because $\varphi(x_j, y_j) > 0$, Cor. 10.23 tells us that for some index i_j, we have

$$\varphi(x_j, y_j) = g^{i_j}(x_j) - y_j^{i_j}, \quad \langle Dg^{i_j}(x_j), \bar{v} \rangle \geqslant -1/j.$$

By taking an appropriate subsequence, we may suppose $i_j \to i_0$; it follows that $\varphi(\bar{x},0) = g^{i_0}(\bar{x})$, whence $i_0 \in I(\bar{x}) \neq \emptyset$. We also obtain

$$\langle Dg^{i_0}(\bar{x}), \bar{v} \rangle \geqslant 0,$$

which contradicts the defining property of $\bar{v}$. □

10.26 Exercise. We return to the robust optimization problem of Example 10.1, which leads to the following special case of problem (P):

$$\text{Minimize } f(x) \text{ subject to } g(x) := \max_{q \in Q} e(x,q) - E \leqslant 0.$$

We shall suppose that x lives in $\mathbb{R}^n$, and that Q is a compact metric space. We further assume that f is continuously differentiable, and that the derivative e'_x with respect to x exists and is continuous in (x,q).

The goal is to express appropriate necessary conditions for optimality. Later, we shall have at our disposal a multiplier rule that applies directly to the problem. Now, however, we show how to proceed on the strength of a certain reduction device for inequality constraints, together with the use of Danskin's theorem.

Let x_* solve the problem, and assume that the inequality constraint is saturated: $g(x_*) = 0$. (Otherwise, the inequality constraint is locally irrelevant, and we have $f'(x_*) = 0$.) We denote by Q_* the set of points $q \in Q$ such that $e(x_*, q) = E$.

Proposition. *There exist $\eta = 0$ or 1, a finite collection $\{q_i : 1 \leqslant i \leqslant k\}$ of k points in Q_*, for some $k \leqslant n+1$, and $\gamma \in \mathbb{R}^k_+$ such that $(\eta, \gamma) \neq 0$ and*

$$0 = \eta f'(x_*) + \sum_{i=1}^{k} \gamma_i \, e'_x(x_*, q_i).$$

For purposes of the proof, we extend the parameter space Q to $Q \times Z := Q \times \{0,1\}$, and we define

$$h(x,q,z) = \begin{cases} f(x) - f(x_*) & \text{if } z = 0 \\ e(x,q) - E & \text{if } z = 1. \end{cases}$$

We further introduce $F(x) = \max\{h(x,q,z) : (q,z) \in Q \times Z\}$. Then F is locally Lipschitz, as follows from Theorem 10.22.

(a) Prove that $F(x) \geqslant 0 \ \forall x$, and that $F(x_*) = 0$.

(b) Deduce that $0 \in \partial_C F(x_*)$, and use Theorem 10.22 to obtain the stated necessary conditions.

(c) If the set $\{e'_x(x_*, q) : q \in Q_*\}$ is assumed to be positively linearly independent, show that the assertion of the proposition may be strengthened: we may take $\eta = 1$ and reduce the number of points q_i involved to at most n.

(d) Suppose now that Q consists of finitely many points. Then the constraint $g(x) \leqslant 0$ is equivalent to a finite number of inequality constraints $e_i(x) \leqslant E$, and Theorem 9.1 can be applied to the problem. Show that the necessary conditions of the theorem are equivalent to those of the proposition above.

(e) Suppose that f is convex, and that e is convex in x. Prove that any point x_* that satisfies $g(x_*) \leqslant 0$ together with the conclusions of the proposition with $\eta = 1$ is a solution of the problem. □

The gradient formula. The celebrated theorem of Rademacher asserts that if a function $f\colon \mathbb{R}^n \to \mathbb{R}$ is Lipschitz on an open set U, then it is differentiable almost everywhere on U (in the sense of Lebesgue measure). It turns out that the derivative of f can be used to generate its generalized gradient, as depicted in the formula below, one of the most useful computational tools in nonsmooth analysis.

It shows that in $\mathbb{R}^n$, $\partial_C f(x)$ can be generated by the values of $\nabla f(u)$ at nearby points u at which $f'(u)$ exists, and furthermore, that points u belonging to any prescribed set of measure zero can be *ignored* in the construction without changing the result. This latter property of $\partial_C f(x)$ is referred to as being "blind to sets of measure zero."

10.27 Theorem. (Gradient formula) *Let $x \in \mathbb{R}^n$, and let $f\colon \mathbb{R}^n \to \mathbb{R}$ be Lipschitz near x. Let E be any subset of zero measure in $\mathbb{R}^n$, and let E_f be the set of points at which f fails to be differentiable. Then*

$$\partial_C f(x) = \operatorname{co}\Big\{ \lim_{i\to\infty} \nabla f(x_i) : x_i \to x,\ x_i \notin E \cup E_f \Big\}.$$

The meaning of the formula is the following: consider any sequence x_i converging to x while avoiding both E and E_f, and such that the sequence $\nabla f(x_i)$ converges to a limit; then the convex hull of all such limits is $\partial_C f(x)$.[2] The proof is postponed for a moment while we study a simple example.

10.28 Example. We proceed to use the gradient formula in order to calculate $\partial_C f(0,0)$, where the function f on $\mathbb{R}^2$ is given by

$$f(x,y) = \max\big\{ \min[2x+y, x], 2y \big\}.$$

Since the Lipschitz property is preserved by max and min, it follows that f is Lipschitz, and in fact piecewise affine. One calculates (see Fig. 10.2)

$$f(x,y) = \begin{cases} 2x+y & \text{for } (x,y) \in A = \{(x,y) : y \leqslant 2x \text{ and } y \leqslant -x\} \\ x & \text{for } (x,y) \in B = \{(x,y) : y \leqslant x/2 \text{ and } y \geqslant -x\} \\ 2y & \text{for } (x,y) \in C = \{(x,y) : y \geqslant 2x \text{ or } y \geqslant x/2\}. \end{cases}$$

Note that $A \cup B \cup C = \mathbb{R}^2$, and that the boundaries of these three sets form (together) a set E of measure 0. If (x,y) does not lie in E, then f is differentiable at (x,y), and $\nabla f(x,y)$ is one of the points $(2,1)$, $(1,0)$, or $(0,2)$. The gradient formula implies that $\partial_C f(0,0)$ is the triangle obtained as the convex hull of these three points. □

10.29 Exercise. We refer to the function f of the example above.

(a) Prove that f does not admit a local minimum at $(0,0)$, and find a descent direction for f at $(0,0)$ (see Exer. 10.7).

[2] As usual, we identify the dual of $\mathbb{R}^n$ with the space itself; for this reason, we view $\partial_C f(x)$ as a subset of $\mathbb{R}^n$.

(b) Show that f is not convex, and that f is not regular at $(0,0)$. □

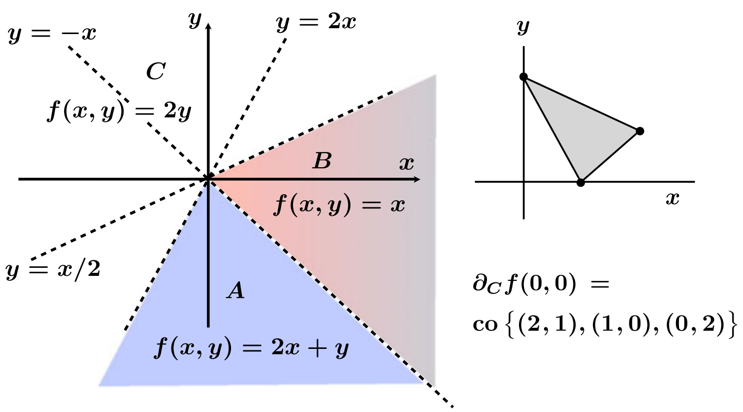


Fig. 10.2 A function f and its generalized gradient at 0

We continue the analysis of f later, in Exer. 10.49.

Proof of Theorem 10.27. Let us note, to begin with, that there is a plenitude of sequences which converge to x and avoid $E \cup E_f$, since the latter has measure 0 near x (by Rademacher's theorem). Since ∇f is locally bounded near x (in view of the Lipschitz condition), one may extract a subsequence for which $\nabla f(x_i)$ converges. It follows that the set on the right in the putative formula is nonempty. We know that $\nabla f(x_i)$ belongs to $\partial_C f(x_i)$ for each i (Exer. 10.4); then the limit of any such sequence must belong to $\partial_C f(x)$ by the closure property of $\partial_C f$ proved in Prop. 10.10. It follows that the set

$$\Big\{ \lim_{i\to\infty} \nabla f(x_i) : x_i \to x,\, x_i \notin E \cup E_f \Big\}$$

is contained in $\partial_C f(x)$. Therefore it is bounded, and in fact compact, since it is closed by its very definition.

Since $\partial_C f(x)$ is convex, we deduce that the left side of the formula asserted by the theorem contains the right. Now, the convex hull of a compact set in $\mathbb{R}^n$ is compact (Exer. 2.8), so to complete the proof, we need only show that the support function of the left side (that is, $f^\circ(x;\cdot)$) never exceeds that of the right. This is what the following lemma does:

Lemma. *For any $v \neq 0$ in $\mathbb{R}^n$, for any $\varepsilon > 0$, we have*

$$f^\circ(x;v) - \varepsilon \leqslant \limsup \big\{ \nabla f(y) \bullet v : y \to x,\, y \notin E \cup E_f \big\}.$$

To prove this, let the right-hand side be denoted α. Then by definition, there is a $\delta > 0$ such that

$$y \in B(x,\delta), \quad y \notin E \cup E_f \implies \nabla f(y) \bullet v \leqslant \alpha + \varepsilon.$$

We also choose δ small enough so that f is Lipschitz on $B(x,\delta)$; thus, $E \cup E_f$ has measure 0 in $B(x,\delta)$. Now consider the line segments

$$L_y = \{ y + t v : 0 < t < \delta/(2|v|) \}.$$

Since $E \cup E_f$ has measure 0 in $x + \delta B$, it follows from Fubini's theorem that for almost every y in $x + (\delta/2)B$, the line segment L_y meets $E \cup E_f$ in a set of 0 one-dimensional measure. Let y be any point in $x + (\delta/2)B$ having this property, and let $0 < t < \delta/(2|v|)$. Then

$$f(y+tv) - f(y) = \int_0^t \nabla f(y+sv) \bullet v \, ds,$$

since f' exists a.e. on L_y. Since we have $\|y+sv-x\| < \delta$ for $0 < s < t$, it follows that $\nabla f(y+sv) \bullet v \leqslant \alpha + \varepsilon$, whence

$$f(y+tv) - f(y) \leqslant t(\alpha + \varepsilon).$$

Since this is true for all y within $\delta/2$ of x except those in a set of measure 0, and for all t in $(0, \delta/(2|v|))$, and since f is continuous, it is in fact true for all such y and t. We deduce $f^{\circ}(x;v) \leqslant \alpha + \varepsilon$, which completes the proof of Theorem 10.27.

10.30 Exercise. Let Ω be a nonempty open bounded subset of $\mathbb{R}^n$. Suppose that the function $g : \overline{\Omega} \to \mathbb{R}$ is Lipschitz, equal to zero on the boundary of Ω, and satisfies $\nabla g = 0$ a.e. in Ω. Prove that g is identically zero. □

10.3 Tangents and normals

Throughout this section, S denotes a nonempty closed subset of a Banach space X. The reader will recall that the **distance function** associated with the set S is defined by

$$d_S(x) = \inf_{y \in S} \|y - x\|,$$

a function which is globally Lipschitz of rank 1. Its generalized gradient will lead to a useful theory of tangents and normals for sets S which are neither classically defined smooth sets, nor convex. The results will encompass these special cases, but others as well.

Exact penalization. The distance function is a useful tool in optimization, for reasons that we proceed to explain. Consider the following constrained optimization problem:

$$\text{Minimize } f(x) \text{ subject to } x \in S. \tag{P}$$

There is a simple yet important technique called *exact penalization* that consists of adding a penalty to the cost function, in order to obtain an equivalence between the original constrained problem and the new penalized, unconstrained one. The following result gives an explicit example of this method in action, using the distance function to express the penalty term.

10.31 Proposition. *Let f be Lipschitz of rank k on an open set U that contains S.*

(a) *If $x_* \in S$ solves* (P), *then, for any $K \geqslant k$, the function $x \mapsto f(x)+K d_S(x)$ attains its* (*unconstrained*) *minimum over U at $x = x_*$.*

(b) *Conversely, suppose that, for some $K > k$, the function $x \mapsto f(x)+K d_S(x)$ attains its minimum over U at $x = x_*$. Then x_* belongs to S and solves* (P).

Proof. Suppose first that $x_* \in S$ solves (P). Let $x \in U$ and $\varepsilon > 0$, and choose $s \in S$ such that $\|x-s\| \leqslant d_S(x)+\varepsilon$. The fact that f is minimized over S at x_*, and the Lipschitz property, yield

$$f(x_*) \leqslant f(s) \leqslant f(x)+k\|s-x\| \leqslant f(x)+k d_S(x)+k\varepsilon.$$

Letting $\varepsilon \downarrow 0$ shows that $f+k d_S$ attains its minimum over U at $x = x_*$. That this also holds when k is replaced by any $K > k$ is evident.

We turn now to the converse. Let the point $x_* \in U$ minimize $f+K d_S$ over U, where $K > k$. We first show that $x_* \in S$, reasoning by the absurd. Suppose to the contrary that $x_* \in U \setminus S$; then $d_S(x_*) > 0$, since S is closed. Pick $s \in S$ so that

$$\|s-x_*\| < (K/k) d_S(x_*)$$

(this is possible because $K > k$). Since f is Lipschitz of rank k, we have

$$f(s) \leqslant f(x_*)+k\|s-x_*\|.$$

Recall that x_* minimizes $x \mapsto f(x)+K d_S(x)$ over U; since $s \in S \subset U$, we have

$$f(x_*)+K d_S(x_*) \leqslant f(s) \leqslant f(x_*)+k\|s-x_*\| < f(x_*)+K d_S(x_*),$$

a contradiction that proves $x_* \in S$. Given this fact, it follows that x_* minimizes f over S, since $d_S = 0$ on S. □

In view of Prop. 10.31, solutions of (P) give rise to critical points of $f+K d_S$:

$$0 \in \partial_C(f+K d_S)(x) \subset \partial_C f(x)+\partial_C d_S(x).$$

This is already a multiplier rule of sorts, one that becomes all the more interesting to the extent that the term $\partial_C d_S(x)$ can be interpreted geometrically. In fact, it turns out that the distance function opens a natural gateway to generalized tangents and normals, as we now see.

10.32 Definition. *Let x be a point in S. The (generalized)* **tangent and normal cones** *to S at x are defined as follows:*

$$T_S^C(x) = \{ v \in X : d_S^\circ(x;v) = 0 \}$$
$$N_S^C(x) = T_S^C(x)^\triangle = \{ \zeta \in X^* : \langle \zeta, v \rangle \leqslant 0 \ \forall v \in T_S^C(x) \}.$$

The reader will observe that we would obtain the same set of tangents by imposing instead the condition $d_S^\circ(x;v) \leqslant 0$, since we always have $d_S^\circ(x;v) \geqslant 0$, as a result of the fact that d_S attains a minimum at x. The definition is a natural one, if we consider that a tangent direction is one in which the distance function will not increase.

It is clear that both definitions above lead to sets which are, in fact, cones. It follows too that when x belongs to int S, then $T_S^C(x) = X$ and $N_S^C(x) = \{0\}$. More generally, two sets which coincide in a neighborhood of x admit the *same* tangent and normal cones at x (since the distance functions agree locally). For this reason, the theory could be developed just as well for sets that are *locally closed* near the point x of interest. (This means that, for some $\delta > 0$, the set $S \cap B(x,\delta)$ is closed.) However, we shall not insist on this refinement.

In Def. 10.32, the normal cone is obtained from the tangent cone by polarity; the reader will recall that this was also the modus operandi adopted earlier in the classical setting (§ 1.4).

10.33 Exercise. In any reasonable concept of normality, we expect normals to products to coincide with products of normals, and similarly for tangents. To be precise, let S_i be a subset of the space X_i, and let $x_i \in S_i$ $(i = 1,2)$. Observe that

$$d_{S_1 \times S_2}(u_1,u_2) = d_{S_1}(u_1) + d_{S_2}(u_2) \ \forall (u_1,u_2) \in X_1 \times X_2.$$

Use this to prove that

$$N^C_{S_1 \times S_2}(x_1,x_2) = N^C_{S_1}(x_1) \times N^C_{S_2}(x_2) \text{ and } T^C_{S_1 \times S_2}(x_1,x_2) = T^C_{S_1}(x_1) \times T^C_{S_2}(x_2).$$

(The analogous facts for the classical tangent and normal cones $T_S(x)$ and $N_S(x)$ have already been noted in Exer. 1.38.) □

It is not always the case that applying polarity to the classical normal cone $N_S(x)$ brings one back to $T_S(x)$. (This must fail when $T_S(x)$ is nonconvex, notably.) However, in the generalized setting, full duality holds, as we now see. (Below, cl* denotes weak* closure.)

10.34 Theorem. *Let $x \in S$. Then*

(a) $T_S^C(x)$ *is a closed convex cone contained in* $T_S(x)$.

(b) $N_S^C(x)$ *is a weak* closed convex cone containing* $N_S(x)$.

(c) *We have* $N_S^C(x) = \text{cl}^*\{\lambda\zeta : \lambda \geqslant 0, \zeta \in \partial_C d_S(x)\}$ and $T_S^C(x) = N_S^C(x)^{\triangle}$.

Proof. In light of Def. 10.3, we have

$$T_S^C(x) = \{ v : \langle \zeta, v \rangle \leqslant 0 \ \forall \zeta \in \partial_C d_S(x) \},$$

which reveals $T_S^C(x)$ as a closed convex cone. Let $v \in T_S^C(x)$. Then $d_S^{\circ}(x;v) \leqslant 0$, so there is a sequence t_i decreasing to 0 such that $\lim_{i\to\infty} d_S(x+t_i v)/t_i = 0$. For each i, let $x_i \in S$ satisfy

$$\| x_i - x - t_i v \| \leqslant d_S(x+t_i v) + t_i^2.$$

Then $v = \lim_{i\to\infty} (x_i - x)/t_i$, which proves that $v \in T_S(x)$. The first part of the theorem is proved.

As a set defined by polarity, $N_S^C(x)$ is automatically a weak* closed convex cone. Since $N_S(x)$ is defined from $T_S(x)$ via polarity, and since taking polars reverses inclusions, we have

$$N_S^C(x) = T_S^C(x)^{\triangle} \supset T_S(x)^{\triangle} = N_S(x),$$

which proves (b).

Let Σ be the set whose closure appears in (c). Let $\lambda_i \zeta_i$ $(i = 1,2)$ be two nonzero points in Σ, and let $t \in [0,1]$. Then

$$(1-t)\lambda_1\zeta_1 + t\lambda_2\zeta_2 = \big[(1-t)\lambda_1 + t\lambda_2\big]\zeta,$$

where

$$\zeta := \frac{(1-t)\lambda_1}{(1-t)\lambda_1 + t\lambda_2}\,\zeta_1 + \frac{t\lambda_2}{(1-t)\lambda_1 + t\lambda_2}\,\zeta_2 \in \partial_C d_S(x).$$

It follows that Σ is a convex cone. Its weak* closure $\overline{\Sigma}$ is a weak* closed convex cone. It is clear that $\partial_C f(x)$, Σ, and $\overline{\Sigma}$ all have the same polar, namely $T_S^C(x)$. By Prop. 4.34, we have

$$\overline{\Sigma} = \overline{\Sigma}^{\triangle\triangle} = T_S^C(x)^{\triangle} = N_S^C(x),$$

which confirms the first assertion of (c). The other one follows from Prop. 4.30, applied to $T_S^C(x)$. □

It is occasionally useful to have the following alternate, direct, characterization of $T_S^C(x)$ on hand, and it is reassuring to know that tangency does not depend on the

choice of equivalent norms for X (as d_S does). The following is to be compared with the characterization of the classical tangent cone $T_S(x)$ given in Exer. 1.37.

10.35 Proposition. *We have $v \in T_S^C(x)$ if and only if, for every sequence x_i in S converging to x and positive sequence t_i decreasing to 0, there exists a sequence v_i converging to v such that $x_i + t_i v_i \in S \ \forall\, i$.*

Proof. Suppose first that $v \in T_S^C(x)$, and that a sequence x_i in S converging to x, along with a sequence t_i decreasing to 0, are given. We must produce the sequence v_i alluded to in the statement of the theorem. Since $d_S^\circ(x;v) = 0$ by assumption, we have

$$\lim_{i\to\infty} \frac{d_S(x_i + t_i v) - d_S(x_i)}{t_i} = \lim_{i\to\infty} \frac{d_S(x_i + t_i v)}{t_i} = 0\,.$$

Let s_i be a point in S which satisfies $\|x_i + t_i v - s_i\| \leqslant d_S(x_i + t_i v) + t_i/i$, and let us set $v_i = (s_i - x_i)/t_i$. Then $v_i \to v$ as a consequence of the above, and we have $x_i + t_i v_i = s_i \in S$, as required.

Now for the converse. Let v have the stated property concerning sequences, and choose a sequence y_i converging to x and t_i decreasing to 0 such that

$$\lim_{i\to\infty} \frac{d_S(y_i + t_i v) - d_S(y_i)}{t_i} = d_S^\circ(x;v)\,.$$

Our purpose is to prove this quantity nonpositive, for then v belongs to $T_S^C(x)$ by definition.

Let s_i in S satisfy $\|s_i - y_i\| \leqslant d_S(y_i) + t_i/i$. It follows that s_i converges to x. Thus there is a sequence v_i converging to v such that $s_i + t_i v_i \in S$. But then, since d_S is Lipschitz of rank 1, we have

$$d_S(y_i + t_i v) \leqslant d_S(s_i + t_i v_i) + \|y_i - s_i\| + t_i \|v - v_i\| \leqslant d_S(y_i) + t_i(\|v - v_i\| + 1/i)\,.$$

We deduce that the limit above is nonpositive, which completes the proof. □

The following result continues the theme of exact penalization, and contains the basic idea by which it induces multiplier rules.

10.36 Proposition. *Let $f : X \to \mathbb{R}$ be Lipschitz of rank k near x, and suppose that x_* minimizes f over S. Then*

$$0 \in \partial_C (f + k\, d_S)(x_*) \subset \partial_C f(x_*) + k\, \partial_C\, d_S(x_*) \subset \partial_C f(x_*) + N_S^C(x_*)\,.$$

Proof. We may suppose that S is contained in an open set U on which f is Lipschitz of rank k; otherwise, just replace S by $S \cap B(0,\varepsilon)$, which affects neither the hypotheses nor the conclusion. According to Prop. 10.31, x_* locally minimizes $f + k\,d_S$. Then $0 \in \partial_C(f + k\,d_S)(x_*)$ by Fermat's rule; the other inclusions follow from the sum rule 10.13 and Theorem 10.34. □

10.37 Exercise. Optimization problems sometimes involve multiple criteria. Consider the case in which it is of interest to minimize *both* f_1 and f_2 relative to a set A, where the functions are continuously differentiable and A is closed.

A point x_* is *Pareto optimal* if there is no $x \in A$ that satisfies both $f_1(x) < f_1(x_*)$ and $f_2(x) < f_2(x_*)$. We proceed to derive a necessary condition for a point to be optimal in this sense.

(a) Show that a Pareto optimal point x_* minimizes over A the function

$$f(x) = \max\big\{ f_1(x) - f_1(x_*),\, f_2(x) - f_2(x_*) \big\}.$$

(b) Deduce the existence of $t \in [0,1]$ such that

$$0 \in (1-t) f_1{}'(x_*) + t f_2{}'(x_*) + N_A^C(x_*). \qquad \square$$

Regular sets. Prop. 10.36 becomes more useful when we are able to interpret the normal cone N_S^C appropriately in various special cases. The rest of this section is devoted to such results. A useful element in this undertaking is the following concept, which extends regularity from functions to sets.

The set S is said to be **regular** at a point $x \in S$ provided that $T_S(x) = T_S^C(x)$. Note that when S is regular at x, we also have, by polarity:

$$N_S^C(x) = T_S^C(x)^{\triangle} = T_S(x)^{\triangle} = N_S(x).$$

In fact, this provides an alternate, equivalent way to characterize regularity:

10.38 Proposition. *S is regular at x if and only if $N_S^C(x) = N_S(x)$.*

Proof. We have already observed the necessity. For the sufficiency, suppose that $N_S^C(x) = N_S(x)$. Then, taking polars,

$$T_S^C(x) = N_S^C(x)^{\triangle} = N_S(x)^{\triangle} = T_S(x)^{\triangle\triangle} \supset T_S(x) \supset T_S^C(x).$$

It follows that we have equality throughout, and that S is regular at x. $\square$

We shall see below that convex sets, as well as sets defined by smooth (nondegenerate) equalities and inequalities (such as manifolds with or without boundary), are regular, and that the new tangents and normals defined above coincide with the familiar ones.

10.39 Theorem. *Let $x \in S$, where S is closed and convex. Then S is regular at x, and we have*

$$T_S^C(x) = T_S(x) = \operatorname{cl}\Big\{ \frac{u-x}{t} : u \in S,\ t > 0 \Big\}$$

$$N_S^C(x) = N_S(x) = \big\{ \zeta \in X^* : \langle \zeta, u-x \rangle \leqslant 0 \ \forall\, u \in S \big\}.$$

Proof. We recall that the given characterizations of T_S and N_S are known results (see Prop. 2.9). To prove the theorem, it suffices to prove $N_S^C(x) \subset N_S(x)$, for then the two normal cones coincide (by Theorem 10.34), and regularity follows from Prop. 10.38.

To prove $N_S^C(x) \subset N_S(x)$, it suffices to show that

$$\zeta \in \partial_C d_S(x) \Longrightarrow \zeta \in N_S(x).$$

Now by Theorem 10.8, such a ζ belongs to $\partial d_S(x)$, since the distance function of a convex set is convex. We may write, therefore, the subgradient inequality:

$$d_S(u) - d_S(x) \geqslant \langle \zeta, u - x \rangle \;\; \forall u \in X,$$

which implies $\langle \zeta, u-x \rangle \leqslant 0 \;\; \forall u \in S$, and thus $\zeta \in N_S(x)$. □

10.40 Exercise. We develop here a geometrical characterization of the generalized normal cone in finite dimensions. S is a nonempty closed subset of $\mathbb{R}^n$, and $\operatorname{proj}_S(x)$ denotes as usual the (nonempty) set of points $u \in S$ satisfying $d_S(x) = |x - u|$.

(a) Prove that $0 \in \partial_C d_S(x) \;\; \forall x \in S$.

(b) Use Prop. 10.36 to prove that if $x \notin S$ and $y \in \operatorname{proj}_S(x)$, then

$$\frac{x-y}{|x-y|} \in \partial_C d_S(y).$$

(These "perpendiculars" to the set S at y generate the *proximal normal cone* to S at y, an object that we shall encounter later.)

(c) Prove that if the derivative $d_S'(x)$ exists at a point $x \in S$, then $d_S'(x) = 0$.

(d) Suppose that $x \notin S$ and that $d_S'(x)$ exists. Let $y \in \operatorname{proj}_S(x)$ and $v \in \mathbb{R}^n$. Show that

$$\lim_{t \downarrow 0} \frac{|x+tv-y| - |x-y|}{t} \geqslant \nabla d_S(x) \bullet v.$$

Deduce that $\nabla d_S(x) = (x-y)/|x-y|$, and that y is the unique point in $\operatorname{proj}_S(x)$.

(e) Let $x \in \partial S$. Use the gradient formula (Theorem 10.27), together with the above, to obtain the following characterization of $\partial_C d_S(x)$:

$$\partial_C d_S(x) = \operatorname{co}\Big\{0, \lim_{i \to \infty} \frac{x_i - y_i}{|x_i - y_i|} : x_i \notin S,\, x_i \to x,\, y_i \in \operatorname{proj}_S(x_i)\Big\}.$$

(f) Let $x \in \partial S$. Prove the following characterization of the generalized normal cone:

$$N_S^C(x) = \overline{\operatorname{co}}\Big\{\lambda \lim_{i \to \infty} \frac{x_i - y_i}{|x_i - y_i|} : \lambda \geqslant 0,\, x_i \notin S,\, x_i \to x,\, y_i \in \operatorname{proj}_S(x_i)\Big\}.$$ □

10.41 Exercise. Let S be the following subset of $\mathbb{R}^2$:

$$S = \{(x,y) : x \leqslant 0\} \cup \{(x,y) : y \leqslant 0\}.$$

We have seen in §1.4 (see Fig. 1.2 (b), p. 25) that the classical tangent cone $T_S(0,0)$ coincides with S, and that $N_S(0,0)$ reduces to $\{(0,0)\}$.

(a) Using Exer. 10.40, prove that $N_S^C(0,0) = \{(x,y) : x \geqslant 0,\ y \geqslant 0\}$.

(b) Deduce that $T_S^C(0,0) = \{(x,y) : x \leqslant 0,\ y \leqslant 0\}$. □

The exercise illustrates the general fact that (relative to the classical constructs) generalized tangent vectors are more stringent, and generalized normal vectors more permissive.

Level and sublevel sets. The sets we meet in practice are often defined by functional relations, notably by equalities and inequalities.

10.42 Theorem. *Let $f : X \to \mathbb{R}$ be a locally Lipschitz function, and define*

$$S = \{u \in X : f(u) \leqslant 0\}.$$

If the point $x \in S$ satisfies $0 \notin \partial_C f(x)$, then we have

$$T_S^C(x) \supset \{v \in X : f^\circ(x;v) \leqslant 0\} \text{ and } N_S^C(x) \subset \{\lambda\zeta : \lambda \geqslant 0,\ \zeta \in \partial_C f(x)\}.$$

If, in addition, f is regular at x, then equality holds in both estimates, S is regular at x, and we have $T_S^C(x) = T_S(x)$, $N_S^C(x) = N_S(x)$.

Proof.

A. We begin with the inclusion involving $T_S^C(x)$. We observe first that there exists v_0 such that $f^\circ(x;v_0) < 0$, for otherwise (since $f^\circ(x;\cdot)$ is the support function of $\partial_C f(x)$) we would have $0 \in \partial_C f(x)$, contrary to hypothesis. If v belongs to the set

$$D := \{v \in X : f^\circ(x;v) \leqslant 0\},$$

then, for any $\varepsilon > 0$, we have $f^\circ(x;v+\varepsilon v_0) < 0$, by subadditivity. Since $T_S^C(x)$ is closed, we see, therefore, that it suffices to prove the inclusion only for points v satisfying $f^\circ(x;v) < 0$.

For such a v, it follows from the definition of $f^\circ(x;v)$ that, for certain positive numbers ε and δ, we have

$$f(y+tv) - f(y) \leqslant -\delta t \quad \forall y \in B(x,\varepsilon),\ t \in (0,\varepsilon).$$

We shall prove that $v \in T_S^C(x)$ by means of the characterization furnished by Prop. 10.35. Accordingly, let x_i be any sequence in S converging to x, and t_i any sequence

decreasing to 0. By definition of S, we have $f(x_i) \leqslant 0$. For all i sufficiently large, we also have

$$f(x_i + t_i v) \leqslant f(x_i) - \delta t_i \leqslant -\delta t_i .$$

It follows that $x_i + t_i v \in S$, which confirms that $v \in T_S^C(x)$.

B. We examine next the inclusion involving $N_S^C(x)$. Consider the convex cone

$$K = \{ \lambda\zeta : \lambda \geqslant 0,\ \zeta \in \partial_C f(x) \},$$

whose polar is $\{ v \in X : f^\circ(x;v) \leqslant 0 \}$. If K is weak* closed, then $K^{\triangle\triangle} = K$ by Prop. 4.34, and the estimate for $N_S^C(x)$ follows by polarity from that for $T_S^C(x)$, as follows:

$$N_S^C(x) = T_S^C(x)^\triangle \subset \{ v \in X : f^\circ(x;v) \leqslant 0 \}^\triangle = K^{\triangle\triangle} = K.$$

The following general result supplies the required fact (take $\Sigma = \partial_C f(x)$).

Lemma. *Let Σ be a weak* compact convex set not containing 0. Then $\mathbb{R}_+\Sigma$ is weak* closed.*

Proof. The definition $\|\sigma\|_* = \sup\{ \langle \sigma, u \rangle : u \in B \}$ reveals that the norm is weak* lsc, as an upper envelope of such functions. It follows from compactness that $\|\sigma\|_*$ is strictly bounded away from 0 for $\sigma \in \Sigma$.

It is easily seen that the set $\mathbb{R}_+\Sigma$ is a convex cone. Thus, by Exer. 8.48, it suffices to prove that the set

$$\Sigma_B = B_* \cap \mathbb{R}_+\Sigma = \{ t\sigma : \sigma \in \Sigma,\ t \geqslant 0,\ \|t\sigma\| \leqslant 1 \}$$

is weak* closed.

Fix $u \in X$. When X^* is equipped with the weak* topology, the map f_u from $\mathbb{R}\times X^*$ to $\mathbb{R}$ defined by $f_u(t,\sigma) = \langle t\sigma, u \rangle$ is continuous (Theorem 3.1). This observation reveals that the set

$$\Gamma = \{ (t,\sigma) \in \mathbb{R}_+ \times \Sigma : f_u(t,\sigma) \leqslant 1\ \ \forall u \in B \} = \{ (t,\sigma) \in \mathbb{R}_+ \times \Sigma : \|t\sigma\| \leqslant 1 \}$$

is weak* closed. It is also bounded, as a consequence of the fact that Σ is weak* compact, and because $\|\sigma\|_*$ is bounded away from 0 on Σ. Thus Γ is weak* compact (Cor. 3.15). But Σ_B is the image of Γ under the continuous map $(t,\sigma) \mapsto t\sigma$. Thus Σ_B is weak* compact, and hence weak* closed. □

C. We now suppose that f is regular at x. Let $v \in T_S(x)$. By definition, there exist sequences x_i in S and $t_i \downarrow 0$ such that $v_i := (x_i - x)/t_i \to v$. Then

$$\begin{aligned} f^\circ(x;v) = f'(x;v) &= \lim_{i\to\infty} \frac{f(x+t_i v) - f(x)}{t_i} = \lim_{i\to\infty} \frac{f(x+t_i v_i) - f(x)}{t_i} \\ &= \lim_{i\to\infty} \frac{f(x_i) - f(x)}{t_i} \leqslant 0, \end{aligned}$$

since $f(x) = 0$ and $f(x_i) \leqslant 0$. By part (A), we deduce $T_S(x) \subset T_S^C(x)$. Since we always have $T_S^C \subset T_S$, this implies the equality of the two tangent cones, as well as equality in the initial estimate for $T_S^C(x)$.

We deduce from this the corresponding equalities involving the normal cones:

$$\begin{aligned} N_S(x) &= T_S(x)^\triangle = \big\{ v : f^\circ(x;v) \leqslant 0 \big\}^\triangle = \big[\partial_C f(x)^\triangle \big]^\triangle \\ &= K^{\triangle\triangle} = K \supset N_S^C(x) \supset N_S(x). \end{aligned}$$

(Note that K was defined in part (B).) □

Remark. The *nondegeneracy* hypothesis (or constraint qualification) $0 \notin \partial_C f(x)$ is essential in Theorem 10.42. The reader may show this with an example in $\mathbb{R}$ (take $f(u) = u^2$).

10.43 Exercise. Let $S \subset X$ be closed, bounded, and convex, with nonempty interior. Prove the existence of a function $\varphi : X \to \mathbb{R}$, convex and Lipschitz, such that

$$S = \big\{ x \in X : \varphi(x) \leqslant 0 \big\} \text{ and } x \in \partial S \implies 0 \notin \partial \varphi(x).$$ □

Manifolds with boundary. The preceding exercise shows that the closure of a convex body is a Lipschitz manifold with boundary. We study next an important class of sets, used in differential geometry and elsewhere, that can be described by a finite number of inequalities as follows:

$$S = \big\{ u \in X : f_i(u) \leqslant 0,\ i = 1, 2, \ldots, k \big\},$$

where each function $f_i : X \to \mathbb{R}$ is C^1 (locally, at least). To avoid degeneracy, it is usually assumed that at the point $x \in S$ of interest, the relevant vectors $f_i'(x)$ are linearly independent (we shall improve upon this). The word "relevant" here is taken to refer to the set $I(x)$ of indices i for which $f_i(x) = 0$. (In the absence of any such indices, x lies in the interior of S, and there is no local geometry to elucidate.)

We shall derive the following as a consequence of Theorem 10.42.

10.44 Corollary. *Let $x \in S$, where S is given as above, where $I(x)$ is nonempty, and where the set $\{ f_i'(x) : i \in I(x) \}$ is positively linearly independent. Then*

$$T_S^C(x) = T_S(x) = \big\{ v \in X : \langle f_i'(x), v \rangle \leqslant 0,\ i \in I(x) \big\},$$

$$N_S^C(x) = N_S(x) = \Big\{ \textstyle\sum_{i \in I(x)} \lambda_i f_i'(x) : \lambda_i \geqslant 0 \Big\}.$$

When the hypothesis above holds, we say that "the active constraints are positively linear independent." The reader may observe that, even in $\mathbb{R}^2$ (for example), in contrast to linear independence, any number of vectors can be positively linear independent.

Proof. We set $f(u) = \max_{i=1,2,\dots,k} f_i(u)$, a function which is Lipschitz near x (since each f_i is). Then S is the set described in Theorem 10.42. Furthermore, f is regular, and we have

$$\partial_C f(x) = \operatorname{co}\{ f_i'(x) : i \in I(x) \},$$

by Danskin's theorem 10.22. We see that the hypothesis of positive linear independence is precisely the same as requiring $0 \notin \partial_C f(x)$. Thus, Theorem 10.42 applies; the result follows. □

Banach manifolds. We return now to the context of Theorem 5.35, that of a set S defined as the solution set of an equation $F(u) = 0$. We shall verify that, under the appropriate nondegeneracy hypothesis, the classical and generalized tangent and normal cones coincide.

10.45 Theorem. *Let x belong to the set $S = \{u \in X : F(u) = 0\}$, where F is continuously differentiable in a neighborhood of x and $F'(x)$ is surjective. Then*

$$T_S(x) = T_S^C(x) = \{ v \in X : F'(x)\,v = 0 \} \text{ and}$$
$$N_S(x) = N_S^C(x) = \{ \zeta = \Lambda \circ F'(x) \in X^* : \Lambda \in Y^* \} = F'(x)^* Y^*.$$

Proof. It suffices to show that any point v that satisfies $F'(x)v = 0$ belongs to $T_S^C(x)$. For then we derive $T_S(x) \subset T_S^C(x)$, in view of Theorem 5.35, whence equality holds. (The equality of the normal cones then follows from taking polars.)

Accordingly, let $F'(x)v = 0$. It follows (exercise) that we have

$$\lim_{y \to x,\, t \downarrow 0} (F(y+tv) - F(y))/t = 0. \tag{1}$$

By Theorem 5.32, there exists K and a neighborhood W of x in which

$$d\big(u, F^{-1}(0)\big) \leqslant K \| F(u) \|_Y \quad \forall u \in W.$$

Let $y_i \to x$ and $t_i \downarrow 0$ be sequences realizing $d_S^\circ(x;v)$; then choose $z_i \in S$ to satisfy $\| y_i - z_i \| < d_S(y_i) + t_i^2$; thus $F(z_i) = 0$. We have, for all i sufficiently large,

$$d_S(z_i + t_i v) = d\big(z_i + t_i v, F^{-1}(0)\big) \leqslant K \| F(z_i + t_i v) - F(z_i) \|_Y .$$

This leads to

$$d_S^\circ(x;v) = \lim_{i \to \infty} \frac{d_S(y_i + t_i v) - d_S(y_i)}{t_i} \leqslant \limsup_{i \to \infty} \frac{d_S(y_i + t_i v) - \| y_i - z_i \| + t_i^2}{t_i}.$$

Because d_S is Lipschitz of rank 1, the last expression is bounded above by

$$\limsup_{i \to \infty} \frac{d_S(z_i + t_i v)}{t_i} \leqslant \limsup_{i \to \infty} \frac{K \| F(z_i + t_i v) - F(z_i) \|_Y}{t_i} = 0,$$

in light of (1). It follows that $d_S^{\circ}(x;v) \leqslant 0$; that is, $v \in T_S^C(x)$. □

We now examine how the characterizations of normal cones to level and sublevel sets that we have proved above contribute to proving and extending the multiplier rule.

10.4 A nonsmooth multiplier rule

We have seen in Prop. 10.36 that when f has a local minimum at x_* relative to S, then the following necessary condition holds:

$$0 \in \partial_C f(x_*) + N_S^C(x_*).$$

If the set S is a level or sublevel set, we can exploit the corresponding characterization of N_S^C to translate this into explicit multiplier form. For example, when Theorem 10.45 is invoked in connection with the necessary condition above, we obtain the following:

10.46 Proposition. *Let x_* minimize f relative to the set $S = \{x \in X : F(x) = 0\}$, where f is continuously differentiable, $F : X \to Y$ is a continuously differentiable mapping to another Banach space Y, and $F'(x_*)$ is surjective. Then there exists $\Lambda \in Y^*$ such that $\{f + \Lambda \circ F\}'(x_*) = 0$.*

The reader will note that, in contrast to Theorem 9.1, the equality constraint here is (potentially) infinite dimensional. We now prove a version of the multiplier rule that conserves this advance and subsumes the smooth case (Theorem 9.1), while allowing the data to be nonsmooth (and nonconvex). Thus, consider anew the problem

$$\text{Minimize } f(x) \text{ subject to } g(x) \leqslant 0, \ h(x) = 0, \ x \in S \tag{P}$$

where, as before, the functions f, g, and h map X to $\mathbb{R}$, $\mathbb{R}^m$, $\mathbb{R}^n$ respectively. We assume that S is closed.

10.47 Theorem. *Let x_* be a solution of* (P), *where f, g, and h are Lipschitz near x_*. Then there exists $(\eta, \gamma, \lambda) \in \mathbb{R} \times \mathbb{R}^m \times \mathbb{R}^n$ satisfying the* **nontriviality** *condition*

$$(\eta, \gamma, \lambda) \neq 0,$$

the **positivity** *and* **complementary slackness** *conditions*

$$\eta = 0 \text{ or } 1, \quad \gamma \geqslant 0, \quad \langle \gamma, g(x_*) \rangle = 0,$$

and the **stationarity** *condition*

$$0 \in \partial_C \{ \eta f + \langle \gamma, g \rangle + \langle \lambda, h \rangle \}(x_*) + N_S^C(x_*).$$

This theorem differs from the classical case (Theorem 9.1), in permitting x_* to lie in the *boundary* of S. When this is the case, the normal cone $N_S^C(x_*)$ provides an additional term that (necessarily) figures in the new stationarity condition. The latter reduces to the earlier one:

$$\{\,\eta f + \langle \gamma, g \rangle + \langle \lambda, h \rangle\}'(x_*) = 0$$

when $x_* \in \operatorname{int} S$, and when the functions involved are smooth. Thus, the proof below is simultaneously a proof of Theorem 9.1.

Proof. Without loss of generality, we may suppose that the functions f, g, h are globally Lipschitz on a neighborhood of the set S, for when $S \cap B(x_*, \delta)$ replaces S (for an arbitrarily small $\delta > 0$), neither the conclusions nor the hypotheses of the theorem are affected.

A. We define, for a fixed $\varepsilon \in (0,1)$:

$$M = \{\, \mu = (\eta, \gamma, \lambda) \in \mathbb{R}_+ \times \mathbb{R}_+^m \times \mathbb{R}^n : |(\eta, \gamma, \lambda)| = 1\,\}$$
$$G_\varepsilon(x) = \max_{\mu \in M} \, (\eta, \gamma, \lambda) \bullet \big(f(x) - f(x_*) + \varepsilon, \, g(x), \, h(x) \big).$$

We claim that $G(x) > 0 \;\; \forall x \in S$. If, to the contrary, $G(x) \leqslant 0$, then it follows from this, in view of the way G_ε is defined, that

$$g(x) \leqslant 0, \; h(x) = 0, \; f(x) \leqslant f(x_*) - \varepsilon,$$

contradicting the optimality of x_* for (P). We also observe $G(x_*) = \varepsilon$. It follows that $G(x_*) \leqslant \inf_{x \in S} G(x) + \varepsilon$. By Theorem 5.19 with $\lambda = \sqrt{\varepsilon}$ and $E = S$, we deduce the existence of $x_\varepsilon \in S$ such that $\|x_\varepsilon - x_*\| \leqslant \sqrt{\varepsilon}$, and

$$\min_{x \in S} \; G(x) + \sqrt{\varepsilon}\, \|x - x_\varepsilon\| = G(x_\varepsilon).$$

Since (f, g, h), and therefore G, is globally Lipschitz in a neighborhood of S, there exists by Theorem 10.31 a number K such that the function

$$H(x) = G(x) + \sqrt{\varepsilon}\, \|x - x_\varepsilon\| + K d_S(x)$$

attains a local minimum at x_ε. Notice that K depends only on a Lipschitz constant for $G(x) + \sqrt{\varepsilon}\, \|x - x_\varepsilon\|$, and may therefore be taken to be independent of ε.

The presence of a local minimum, together with nonsmooth calculus (see Prop. 10.5 and Theorem 10.13) implies

$$0 \in \partial_C H(x_\varepsilon) \subset \partial_C G(x_\varepsilon) + \sqrt{\varepsilon}\, B_* + K \partial_C d_S(x_\varepsilon).$$

B. The next result examines a certain max function that helps us to interpret the last inclusion above.

Lemma. *Let $\varphi : X \to \mathbb{R}^k$ be Lipschitz near x. Let M be a compact set in $\mathbb{R}^k$, and define $G : X \to \mathbb{R}$ by*

$$G(x) = \max_{\mu \in M} \mu \bullet \varphi(x).$$

Then G is Lipschitz near x. For any u near x, we set

$$M(u) = \{\mu \in M : G(u) = \mu \bullet \varphi(u)\},\ u \in X.$$

Then, if $M(x)$ is a singleton $\{\mu_0\}$, we have

$$\partial_C G(x) \subset \partial_C(\mu_0 \bullet \varphi)(x).$$

The fact that G is Lipschitz near x is left to the reader to show. To prove the stated inclusion, it suffices to prove that, for all $v \in X$, we have

$$G^\circ(x;v) \leqslant (\mu_0 \bullet \varphi)^\circ(x;v).$$

Let $y_i \to x$ and $t_i \downarrow 0$ be sequences realizing $G^\circ(x;v)$, and let $\mu_i \in M(y_i + t_i v)$. Taking a subsequence if necessary, we may suppose that μ_i converges to a limit $\bar{\mu}$. We see easily that $\bar{\mu} \in M(x)$, whence $\bar{\mu} = \mu_0$. Then $G^\circ(x;v)$ is given by

$$\begin{aligned}
&\lim_{i\to\infty} \frac{G(y_i + t_i v) - G(y_i)}{t_i} \leqslant \limsup_{i\to\infty} \frac{\mu_i \bullet \varphi(y_i + t_i v) - \mu_i \bullet \varphi(y_i)}{t_i} \\
&\leqslant \limsup_{i\to\infty} \frac{\mu_0 \bullet \varphi(y_i + t_i v) - \mu_0 \bullet \varphi(y_i)}{t_i} + \limsup_{i\to\infty} |\mu_i - \mu_0| \frac{|\varphi(y_i + t_i v) - \varphi(y_i)|}{t_i} \\
&\leqslant (\mu_0 \bullet \varphi)^\circ(x;v) + \limsup_{i\to\infty} K|\mu_i - \mu_0|\,\|v\| \quad (K \text{ is a Lipschitz constant for } \varphi) \\
&= (\mu_0 \bullet \varphi)^\circ(x;v),
\end{aligned}$$

proving the lemma.

C. We now show that there is a unique point $\mu^\varepsilon = (\eta^\varepsilon, \gamma^\varepsilon, \lambda^\varepsilon)$ in $M(x_\varepsilon)$, the subset of M at which the maximum defining $G(x_\varepsilon)$ is attained. To see this, suppose that $(\eta^i, \gamma^i, \lambda^i)$ $(i = 1, 2)$ are two distinct such points. Then there exists $t > 1$ such that the point

$$(\eta, \gamma, \lambda) = \frac{t}{2}\,(\eta^1, \gamma^1, \lambda^1) + \frac{t}{2}\,(\eta^2, \gamma^2, \lambda^2)$$

belongs to M. But then

$$G(x_\varepsilon) \geqslant (\eta, \gamma, \lambda) \bullet \big(f(x_\varepsilon) - f(x_*) + \varepsilon\,, g(x_\varepsilon), h(x_\varepsilon)\big) = t\,G(x_\varepsilon) > G(x_\varepsilon),$$

since $G(x_\varepsilon) > 0$. This contradiction shows that $M(x_\varepsilon)$ is a singleton $(\eta^\varepsilon, \gamma^\varepsilon, \lambda^\varepsilon)$.

We may therefore invoke the lemma to deduce the existence of $\zeta_\varepsilon \in X^*$ with $\|\zeta_\varepsilon\|_* \leqslant \sqrt{\varepsilon}$ such that

$$\zeta_\varepsilon \in \partial_C \big\{ (\eta^\varepsilon, \gamma^\varepsilon, \lambda^\varepsilon) \bullet (f, g, h) \big\}(x_\varepsilon) + K \partial_C\, d_S(x_\varepsilon). \qquad (1)$$

D. We obtain the conclusion (1) for a sequence $\varepsilon_i \downarrow 0$. We denote by $x_i, (\eta^i, \gamma^i, \lambda^i)$ and ζ_i the corresponding sequences $x_\varepsilon, (\eta^\varepsilon, \gamma^\varepsilon, \lambda^\varepsilon)$ and ζ_ε. We may suppose, by taking a subsequence, that $(\eta^i, \gamma^i, \lambda^i)$ converges to a limit $(\eta, \gamma, \lambda) \in M$. Of course, we have $x_i \to x_*$, $\zeta_i \to 0$. It follows from (1) and the sum rule (Theorem 10.13) that

$$\zeta_i \in \partial_C \big\{ (\eta, \gamma, \lambda) \bullet (f, g, h) \big\}(x_i) + K \partial_C d_S(x_i) + \big|(\eta^i, \gamma^i, \lambda^i) - (\eta, \gamma, \lambda)\big| K B_* .$$

Passing to the limit (and using Prop. 10.10), we find

$$0 \in \partial_C \big\{ (\eta, \gamma, \lambda) \bullet (f, g, h) \big\}(x_*) + K \partial_C d_S(x_*),$$

which yields the desired stationarity condition, since $\partial_C d_S(x_*)$ is contained in $N_S^C(x_*)$.

Suppose now that $g^j(x_*) < 0$ for a certain component g^j of g. Then for i sufficiently large, we have $g^j(x_i) < 0$. The fact that $\mu_i := (\eta^i, \gamma^i, \lambda^i)$ maximizes

$$\mu \bullet \big(f(x_i) - f(x_*) + \varepsilon_i, g(x_i), h(x_i) \big) \text{ subject to } \mu \in M,$$

and that the maximum is strictly positive, implies that $(\gamma^i)^j = 0$. We obtain in the limit $\gamma^j = 0$. Thus we derive the complementary slackness condition:

$$\langle \gamma, g(x_*) \rangle = \sum_j \gamma^j g^j(x_*) = 0.$$

If $\eta = 0$, the multiplier $(0, \gamma, \lambda)$ satisfies all the required conclusions. If $\eta > 0$, we may replace (η, γ, λ) by $\big(1, \gamma/\eta, \lambda/\eta\big)$ to obtain the normalized (and normal) multiplier. □

10.48 Exercise. We consider the problem

$$\text{Minimize } f(x) \text{ subject to } \varphi(x) \in \Phi, \tag{Q}$$

where Φ is a closed subset of $\mathbb{R}^k$. We suppose that $f : X \to \mathbb{R}$ and $\varphi : X \to \mathbb{R}^k$ are locally Lipschitz. The problem (Q) may appear to be more general than our standard problem (P), since the latter evidently corresponds to the case in which

$$\varphi(x) = \big(g(x), h(x), x\big) \in \Phi := \mathbb{R}^n_- \times \{0\} \times S.$$

Show, however, that (Q) can in turn be obtained as a special case of (P). Thus, the two abstract problems are equivalent.

We remark that the (Q) form of the optimization problem, though it is classically less familiar, has the advantage of subsuming the cases in which the equality or inequality constraints are absent, without the need to treat them separately.

The appropriate multiplier rule for (Q) is that, for a solution x_*, there exists $\eta = 0$ or 1 and $\nu \in N_\Phi^C(\varphi(x_*))$ such that $(\eta, \nu) \neq 0$ and

$$0 \in \partial_C \big\{ \eta f + \langle \nu, \varphi \rangle \big\}(x_*).$$

Derive this multiplier rule from Theorem 10.47. Conversely, show that Theorem 10.47 follows from it. □

10.49 Example. We consider the problem of minimizing $f(x,y)$ over the points (x,y) in the unit ball: $x^2 + y^2 \leqslant 1$, where f is the function of Example 10.28.

Since the set of admissible points is compact and f is continuous, it is clear that the minimum is attained. We know that f is neither convex nor smooth, so the only multiplier rule that applies here is Theorem 10.47. It yields the stationarity condition

$$(0,0) \in \partial_C \big\{ \eta f + \gamma (x^2 + y^2 - 1) \big\} = \eta\, \partial_C f(x,y) + 2\gamma(x,y),$$

(see Exer. 10.16). If $\eta = 0$, then $\gamma > 0$, which implies that the inequality constraint is saturated: $x^2 + y^2 = 1$. The resulting stationarity condition $(x,y) = (0,0)$ then provides a contradiction. Thus, we may take $\eta = 1$.

Now if $\gamma = 0$, we have $(0,0) \in \partial_C f(x,y)$. However, we claim that $\partial_C f(x,y)$ never contains $(0,0)$. We know this is true at the origin (see Example 10.28); at points in the interior of the zones A, B, or C, $\partial_C f$ is a nonzero singleton. So in proving the claim, we may limit attention to nonzero points on the boundaries between zones.

From the gradient formula, we see that at such points, $\partial_C f$ is a line segment, either $\operatorname{co}\{(2,1),(0,2)\}$ (the A/C boundary) or $\operatorname{co}\{(2,1),(1,0)\}$ (the A/B boundary) or $\operatorname{co}\{(1,0),(0,2)\}$ (the B/C boundary); none of these segments contains $(0,0)$. Conclusion: $\gamma > 0$, and the solution (x,y) lies on the unit circle. In light of this, the stationarity condition reduces to

$$(x,y) \in -1/(2\gamma)\, \partial_C f(x,y).$$

Suppose the solution (x,y) lies in the interior of zone A. Then the stationarity condition affirms that (x,y) is a negative multiple of $(2,1)$; but this is inconsistent with being in A, as the reader may easily verify. A similar argument rules out the interiors of B and C.

There remain three candidates: the points along the inter-zone boundaries lying on the unit circle. For the B/C boundary, this is a point of the form $(2\varepsilon, \varepsilon)$ (for $\varepsilon > 0$), a negative multiple of which would have to lie in the segment $\operatorname{co}\{(1,0),(0,2)\}$, by the stationarity condition; this is impossible, as it is easy to see. The A/B boundary provides a point of the form $(\varepsilon, -\varepsilon)$, a negative multiple of which would have to lie in the segment $\operatorname{co}\{(2,1),(1,0)\}$: impossible. The A/C boundary provides a point of the form $(-\varepsilon, -2\varepsilon)$, a negative multiple of which would lie in the segment $\operatorname{co}\{(2,1),(0,2)\}$. This *is* possible, and it identifies $(x,y) = -(1,2)/\sqrt{5}$ as the only point satisfying the necessary conditions.

Since we know a solution exists and the necessary conditions must hold, we deduce (from the aptly named deductive method) that this is the unique solution. □

Chapter 11

Proximal analysis

We proceed in this chapter to develop the calculus (and the geometry) associated with the proximal subdifferential. The reader will no doubt remember having encountered this object in Chapter 7 (§7.3). As we saw at that time, the utility of the proximal subdifferential is intrinsically limited to cases in which the underlying normed space X is smooth, in a certain sense that is always true of Hilbert spaces. In this chapter, we limit attention to the case $X = \mathbb{R}^n$.

Let us revisit the basic definition, for which we require a function $f : \mathbb{R}^n \to \mathbb{R}_\infty$, and a point $x \in \operatorname{dom} f$. We say that $\zeta \in \mathbb{R}^n$ is a **proximal subgradient** of f at x if for some $\sigma = \sigma(x, \zeta) \geqslant 0$, and for some neighborhood $V = V(x, \zeta)$ of x, we have

$$f(y) - f(x) + \sigma |y - x|^2 \geqslant \langle \zeta, y - x \rangle \ \ \forall y \in V.$$

This is referred to as the *proximal subgradient inequality*. In writing it, we generally prefer the notation of the duality pairing $\langle \zeta, y - x \rangle$ to that of the inner product $\zeta \bullet (y - x)$.

Proximal subgradients (which may be referred to as P-subgradients, or even P-gradients, for short) admit a natural geometrical interpretation, as pointed out in Example 7.28: they correspond to the contact slopes of locally supporting parabolas to the epigraph of f at the point $(x, f(x))$. The collection of all such ζ as described above (which may be empty) constitutes the *proximal subdifferential* of f at x; it is denoted $\partial_P f(x)$.

11.1 Proximal calculus

Here is a first connection between proximal subgradients, generalized gradients, and classical derivatives:

F. Clarke, *Functional Analysis, Calculus of Variations and Optimal Control*, Graduate Texts in Mathematics 264, DOI 10.1007/978-1-4471-4820-3_11,

11.1 Proposition. *Let f be Lipschitz near x. Then $\partial_P f(x) \subset \partial_C f(x)$. If f is C^1 near x and f' is Lipschitz near x (in particular, if f is C^2 near x), then we have*

$$\partial_P f(x) = \{ f'(x) \} = \partial_C f(x).$$

Proof. An immediate consequence of the proximal subgradient inequality is that, for any given v, for all $t > 0$ sufficiently small, we have

$$\big(f(x+tv) - f(x)\big)/t \geqslant \langle \zeta, v \rangle - \sigma t |v|^2.$$

It follows from this, and the definition of $f^{\circ}(x;v)$, that $f^{\circ}(x;v) \geqslant \langle \zeta, v \rangle$, whence ζ belongs to $\partial_C f(x)$ by definition of the generalized gradient.

If f is C^1 near x, then $\partial_C f(x) = \{ f'(x) \}$ by Theorem 10.8, and the argument above yields $\partial_P f(x) \subset \{ f'(x) \}$. The fact that equality holds when f' is Lipschitz near x follows directly from Cor. 7.32. □

It is clear from the definition that for any $k > 0$, we have $\partial_P (kf)(x) = k\, \partial_P f(x)$. In contrast to the generalized gradient, however, this fails when k is negative. For example, as we have seen earlier, when $f(x) = |x|$, we have (see Prop. 7.26 and Example 7.30)

$$\partial_P f(0) = B(0,1), \ \partial_P(-f)(0) = \emptyset.$$

We now begin a systematic study of proximal calculus. The following theorem due to Clarke and Ledyaev is the cornerstone of our development. It is a *multidirectional* extension of the mean value theorem that gives, in certain cases, a more subtle bound on certain increments, where the classic mean value theorem gives an upper bound on all of them. In its statement, the interval notation $[x, Y]$ refers to $\operatorname{co}\{\{x\} \cup Y\}$, the convex hull of the set Y and the point x.

11.2 Theorem. (Mean value inequality) *Let $x \in \operatorname{dom} f$, where $f : \mathbb{R}^n \to \mathbb{R}_\infty$ is lower semicontinuous, and let Y be a compact convex subset of $\mathbb{R}^n$. Then, for any real number $r < \min_Y f - f(x)$ and every $\varepsilon > 0$, there exists $x_* \in [x,Y] + \varepsilon B$ and ζ in $\partial_P f(x_*)$ such that*

$$r \leqslant \langle \zeta, y - x \rangle \ \forall y \in Y.$$

The point x_* may be taken to satisfy

$$f(x_*) \leqslant \min\big\{ f(u) : u \in [x,Y] \big\} + |r|.$$

Proof.

Without loss of generality we take $x = 0$ and $f(0) = 0$; we may also assume that f is globally bounded below by $-m$, for some $m > |r|$. By lower semicontinuity, there exists $\delta \in (0, \varepsilon)$ such that

$$u \in Y + \delta B \implies f(u) > r.$$

We choose $k > 0$ large enough so that

$$y \in Y,\ k\,|\,y - u\,|^2 \leqslant m + r \implies u \in Y + \delta B.$$

We now consider the minimization of the function

$$g(u,y,t) := f(u) + k\,|\,ty - u\,|^2 - rt \quad \text{over} \quad u \in \mathbb{R}^n,\ y \in Y,\ t \in [0,1].$$

The minimum is attained at a point (x_*, y_*, t_*). Since $g(0,y,0) = 0$, we must have $g(x_*, y_*, t_*) \leqslant 0$. We claim that $t_* < 1$. Let us suppose the contrary. Then $g(x_*, y_*, 1) \leqslant 0$. But

$$g(x_*, y_*, 1) = f(x_*) + k\,|\,y_* - x_*\,|^2 - r,$$

so we derive $k\,|\,y_* - x_*\,|^2 \leqslant m + r$, which in turn implies $x_* \in Y + \delta B$ by choice of k, whence $g(x_*, y_*, 1) > 0$ by choice of δ. This contradiction establishes that $t_* \neq 1$.

We now verify the upper bound on $f(x_*)$. Specializing to the points $x = ty$ in the minimization of g yields

$$f(x_*) - rt_* \leqslant g(x_*, y_*, t_*) \leqslant \min\big\{\, f(ty) - rt : y \in Y,\ t \in [0,1]\,\big\},$$

whence

$$f(x_*) \leqslant \min\big\{\, f(ty) + r(t_* - t) : y \in Y,\ t \in [0,1]\,\big\} \leqslant \min_{[0,Y]} f + |\,r\,|.$$

We proceed to write the necessary conditions corresponding to the minimization of g over $\mathbb{R}^n \times Y \times [0,1)$, using Prop. 7.31 (for u) and Prop. 1.39 (for y):

$$\zeta := 2k\,(t_* y_* - x_*) \in \partial_P f(x_*),\ \ 2kt_*(x_* - t_* y_*) = \nu \in N_Y(y_*),\ \ \langle\, \zeta, y_* \,\rangle \geqslant r,$$

with equality in the last relation if $t_* > 0$.

Consider now the case $t_* > 0$. From the above we have $\zeta = -\nu/t_*$, so that, writing the inequality for a normal vector in the sense of convex analysis, we derive

$$\langle\, \zeta,\, y_* - y \,\rangle \leqslant 0 \ \ \forall\, y \in Y.$$

Since we also have $\langle\, \zeta, y_* \,\rangle = r$, it follows that $\langle\, \zeta, y \,\rangle \geqslant r \ \ \forall\, y \in Y$, which is the required conclusion.

There remains the case $t_* = 0$. Then $\zeta = -2kx_*$, and the necessary conditions above hold (with this ζ) for *any* y_* in Y (since the choice of y_* has no effect on the cost). The conditions above then yield $\langle\, \zeta, y_* \,\rangle \geqslant r \ \ \forall\, y_* \in Y$, which, once more, is the required conclusion. □

Note that the theorem does not exclude the case $\min_Y f = \infty$; that is, the case in which $\operatorname{dom} f \cap Y = \emptyset$. A slightly simpler statement is obtained when this possibility is excluded:

11.3 Corollary. *If we add to the hypotheses of* Theorem 11.2 *the condition that* $Y \cap \operatorname{dom} f \neq \emptyset$, *then, for any* $\varepsilon > 0$, *there is a point* $x_* \in [x,Y] + \varepsilon B$ *and* ζ *in* $\partial_P f(x_*)$ *such that*

$$\min_Y f - f(x) \leqslant \langle \zeta, y - x \rangle + \varepsilon \ \ \forall y \in Y, \quad f(x_*) \leqslant \max\big(f(x), \min_Y f\big).$$

Proof. We apply the theorem with $r = \min_Y f - f(x) - \varepsilon$. The new upper bound on $f(x_*)$ follows from the previous one, as may easily be seen by considering separately the two cases $r > 0$ and $r \leqslant 0$. □

11.4 Exercise. In the context of Theorem 11.2, suppose that $\operatorname{dom} f \cap Y = \emptyset$. Prove that for every $\varepsilon > 0$, there exists a point $z \in [x,Y] + \varepsilon B$ and $\zeta \in \partial_P f(z)$ such that $|\zeta| > 1/\varepsilon$. □

The following "uni-directional case" of the theorem, in which Y is a singleton, corresponds to the familiar classical mean value theorem, extended here to lower semicontinuous functions:

11.5 Corollary. *Let* $f : \mathbb{R}^n \to \mathbb{R}_\infty$ *be lsc, and let* x, y *be points in* $\operatorname{dom} f$. *For every* $\varepsilon > 0$, *there exists* $x_* \in [x,y] + \varepsilon B$ *such that, for some* $\zeta \in \partial_P f(x_*)$, *one has*

$$f(y) - f(x) \leqslant \langle \zeta, y - x \rangle + \varepsilon.$$

11.6 Exercise.

(a) Prove Corollary 11.5.

(b) In the context of that corollary, take $n = 2$ and

$$f(x_1, x_2) = -|x_2|, \ x = (-1,0), \ y = (1,0).$$

Show that the point x_* does not lie in the segment $[x,y]$.

(c) Let $\varphi : [a,b] \to \mathbb{R}$ be a continuous function such that

$$t \in (a,b), \, \zeta \in \partial_P \varphi(t) \implies \zeta \leqslant 0.$$

Prove that $\varphi(b) \leqslant \varphi(a)$. □

It is a familiar fact in calculus that a continuously differentiable function f is Lipschitz of rank K if and only if $|\nabla f(x)| \leqslant K \ \ \forall x$. The following consequence of Theorem 11.2 extends that result to lsc functions, and hints at the advantage of

proximal characterizations: one need only check points that are special, in the sense that $\partial_P f(x) \neq \emptyset$.

11.7 Corollary. *Let U be an open convex set in $\mathbb{R}^n$. Let $f : U \to \mathbb{R}_\infty$ be lsc, and let $K \geqslant 0$. Prove that*

$$|f(y) - f(x)| \leqslant K|y - x| \;\forall x, y \in U \iff |\zeta| \leqslant K \;\forall \zeta \in \partial_P f(x) \;\forall x \in U.$$

We now focus for a moment upon the directionality aspect of Theorem 11.2, with a smooth function f for the sake of simplicity.

11.8 Corollary. *Let $f : \mathbb{R}^n \to \mathbb{R}$ be continuously differentiable in a neighborhood of the compact convex subset Y of $\mathbb{R}^n$, and let $x \in Y$. Then there exists $x_* \in Y$ such that $f(x_*) \leqslant f(x)$ and*

$$\min_{y \in Y} \big(f(y) - f(x) \big) \leqslant \min_{y \in Y} \langle f'(x_*), y - x \rangle.$$

Proof. Let ε_i be a positive sequence decreasing to 0, and apply Cor 11.3. There results a point x_i in $Y + \varepsilon_i B$ such that $f(x_i) \leqslant f(x)$ and

$$\min_Y f - f(x) \leqslant \min_{y \in Y} \langle f'(x_i), y - x \rangle + \varepsilon_i.$$

We may suppose (by taking a subsequence) that x_i converges to a limit x_*, a point which is easily seen to satisfy all the stated conditions. □

The classical mean value theorem would provide above an estimate of the form

$$\max_{y \in Y} \big(f(y) - f(x) \big) \leqslant \max_{y \in Y} \max_{z \in [x,y]} \langle f'(z), y - x \rangle,$$

an *upper* bound on the increments of the type that is such a familiar tool in analysis. Cor. 11.8, however, provides a *lower* version of the estimate, with different consequences. Here is an example:

11.9 Exercise. Let $f : \mathbb{R}^n \to \mathbb{R}$ be continuously differentiable. Suppose that, for a certain $\delta > 0$, one has

$$x \in B(0,1) \implies |f'(x)| \geqslant \delta.$$

Invoke Cor. 11.8 to prove that $\min_{B(0,1)} f \leqslant f(0) - \delta$. (Notice that an estimate of the type provided by the traditional mean value theorem does not help here.) □

Fuzzy calculus. Theorem 11.2 illustrates the general situation in proximal calculus, whereby formulas can only be asserted to a given positive tolerance ε. The calculus is said to be *fuzzy*, an adjective that is not meant pejoratively.

The proof of Cor. 11.8 illustrates how to pass to the limit in a fuzzy conclusion in order to obtain a sharp one, when f is smooth. Something similar can be done in a nonsmooth setting as well, by applying a closure operation to the proximal subdifferential, as we now proceed to do.

11.10 Definition. *The* **limiting subdifferential** $\partial_L f(x)$ *is defined as follows*:

$$\partial_L f(x) = \big\{ \zeta = \lim_{i\to\infty} \zeta_i : \zeta_i \in \partial_P f(x_i),\ x_i \to x,\ f(x_i) \to f(x) \big\}.$$

In this definition, we consider all sequences x_i converging to x which admit a corresponding convergent sequence ζ_i, and are such that $f(x_i) \to f(x)$. (The last requirement is, of course, superfluous if f is continuous at x.) It follows from this definition that the L-subdifferential $\partial_L f(x)$ contains the P-subdifferential $\partial_P f(x)$, and that, by its very construction, the graph of the multifunction $x \mapsto \partial_L f(x)$ is closed:

$$\zeta_i \in \partial_L f(x_i),\ x_i \to x,\ \zeta_i \to \zeta \implies \zeta \in \partial_L f(x).$$

We remark that ∂_L, like ∂_P, but unlike ∂_C, is truly a *sub*differential, and that $\partial_L(-f)(x) \neq -\partial_L f(x)$ in general. When we calculate $\partial_L f(x)$ via limits of the form $\lim_{i\to\infty} \zeta_i$, where $\zeta_i \in \partial_P f(x_i)$, no points x_i can be ignored *a priori*, without potentially affecting the result. In particular, and in contrast to the gradient formula (Theorem 10.27), $\partial_L f$ fails to be "blind to sets of measure 0." We illustrate this now.

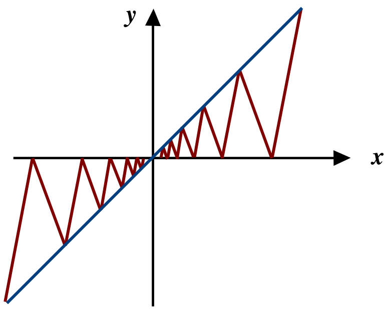


Fig. 11.1 A function f

11.11 Example. We consider a function f whose graph is contained between the x-axis and the graph of $y = x$; f is piecewise affine, with slopes alternating between ± 2; see Fig. 11.1. The function f is clearly Lipschitz, and, by the gradient formula, we find that $\partial_C f(0)$ is the interval $[-2,2]$.

Let E consist of the origin, together with all points at which f is nondifferentiable. Then E is countable, and for $x \notin E$, we have $\partial_P f(x)$ equal to either $\{2\}$ or $\{-2\}$. If we were to ignore the points in E in calculating $\partial_L f(0)$, we would obtain $\{-2,2\}$. But this is incorrect, for at all points $x > 0$ in E for which $f(x) = 0$, we have $\partial_P f(x) = [-2,2]$; thus, $\partial_L f(0) = [-2,2] = \partial_C f(0)$. (We invite the reader to prove that $\partial_P f(0) = \emptyset$.) □

11.12 Proposition. *Let f be Lipschitz near x. Then $\emptyset \neq \partial_L f(x) \subset \partial_C f(x)$. If f is C^1 near x, we have $\partial_L f(x) = \{f'(x)\}$. If f is convex, we have $\partial_L f(x) = \partial f(x)$.*

Proof. According to the proximal density theorem 7.34, there exist points x_i converging to x which admit elements $\zeta_i \in \partial_P f(x_i)$ such that $|f(x_i) - f(x)| < 1/i$. If K is a Lipschitz constant for f in a neighborhood of x, then we have, for all i sufficiently large,

$$\zeta_i \in \partial_P f(x_i) \subset \partial_C f(x_i) \subset B(0,K).$$

Thus, ζ_i is a bounded sequence. By passing to a subsequence, we may suppose that ζ_i converges to a limit ζ, and then we obtain $\zeta \in \partial_L f(x)$ by definition. That $\partial_L f(x)$ is contained in $\partial_C f(x)$ is a consequence of Propositions 10.10 and 11.1. This proves the first assertion.

When f is C^1 near x, we have $\emptyset \neq \partial_L f(x) \subset \partial_C f(x) = \{f'(x)\}$, which implies $\partial_L f(x) = \{f'(x)\}$. If f is convex, and if $\zeta = \lim \zeta_i$ as in the definition of $\partial_L f(x)$, then we have by Prop. 7.26:

$$f(y) - f(x_i) \geqslant \langle \zeta_i, y - x_i \rangle \;\; \forall y \in \mathbb{R}^n.$$

Passing to the limit gives $f(y) - f(x) \geqslant \langle \zeta, y - x \rangle$, whence $\zeta \in \partial f(x)$. It follows that $\partial_L f(x) \subset \partial f(x) = \partial_P f(x) \subset \partial_L f(x)$, whence equality holds. □

11.13 Example. Let us determine $\partial_P f(0,0)$ and $\partial_L f(0,0)$ for the function f of Example 10.28; recall that we have already calculated $\partial_C f(0,0)$. Let (a,b) belong to $\partial_P f(0,0)$. Then, for some $\sigma \geqslant 0$, near $(0,0)$, we have

$$f(x,y) + \sigma\left(|x|^2 + |y|^2\right) \geqslant f(0,0) + ax + by = ax + by.$$

For points (x,y) of the form $(2y,y)$, this becomes $2y + 5\sigma x^2 \geqslant 2ay + by$, for y near 0; this holds if and only if $2a + b = 2$. A similar analysis for points of the form $(x, 2x)$ yields $a + 2b = 4$. These two equations identify a unique possibility: $(a,b) = (0,2)$.

We can show that this is in fact an element of $\partial_P f(0,0)$, by verifying the proximal subgradient inequality in each of the three zones A, B, C. In B, for example, $f(x,y) = x$, so that the proximal inequality requires: $x \geqslant 2y$, which is true in B; the other two follow similarly. We conclude that $\partial_P f(0,0$ is the singleton $\{(0,2)\}$.

In order to calculate $\partial_L f(0,0)$, we need to examine the origin itself (we know that $\partial_P f(0,0) = \{(0,2)\}$), the points interior to the zones A, B, and C (where $\partial_P f$ is a singleton, one of the three gradient values), as well as the boundaries between the zones, since, as we know, they cannot be ignored despite being of measure zero.

Near the boundary between A and B (points of the form $(\varepsilon, -\varepsilon)$, $\varepsilon > 0$), the function f coincides with the function $\min(2x + y, x)$, a concave function which is nondifferentiable at $(\varepsilon, -\varepsilon)$; it follows that $\partial_P f = \emptyset$ at these points. Near the B/C

boundary, f is locally given by max $(x,2y)$, a convex function whose subdifferential (at $(2\varepsilon,\varepsilon)$) is the segment $[(1,0),(0,2)]$. A similar analysis shows that along the A/C boundary, $\partial_P f$ is the segment $[(0,2),(2,1)]$.

We conclude that $\partial_L f(0,0)$ is the (nonconvex) set consisting of these two segments. Note that the convex hull of $\partial_L f(0,0)$ is $\partial_C f(0,0)$, a general fact to be proved later.

The reader will understand from the above that the actual calculation of $\partial_L f$ is generally much more problematic than that of $\partial_C f$. For this reason, most numerical applications of nonsmooth analysis use the generalized gradient, or else limit attention to regular functions f, for which $\partial_L f(x) = \partial_C f(x)$ (see Prop. 11.23). □

Returning now to the mean value inequality, here is a limiting version of Cor. 11.5 which illustrates the role of ∂_L in expressing "non-fuzzy" conclusions:

11.14 Corollary. *Let $f : \mathbb{R}^n \to \mathbb{R}_\infty$ be Lipschitz on a neighborhood of the segment $[x,y]$. Then there exists $x_* \in [x,y]$ such that, for some $\zeta \in \partial_L f(x_*)$, one has*

$$f(y) - f(x) \leqslant \langle \zeta, y - x \rangle.$$

Proof. Let ε_i be a positive sequence decreasing to 0, and apply Cor. 11.5 with $\varepsilon = \varepsilon_i$. The resulting x_{*i} and ζ_i are bounded, so that, by passing to a subsequence, they can be supposed to converge to limits x_* and ζ. Then $\zeta \in \partial_L f(x_*)$, and the result follows. □

11.15 Exercise. If f is Lipschitz near x, and if $\partial_L f(x)$ is a singleton $\{\zeta\}$, then f is differentiable at x, with $\nabla f(x) = \zeta$. □

The following "fuzzy sum rule" is a basic result in proximal calculus.

11.16 Theorem. (Proximal sum rule) *Let $f_1, f_2 : \mathbb{R}^n \to \mathbb{R}_\infty$ be lsc, and let x be a point in $\operatorname{dom} f_1 \cap \operatorname{dom} f_2$. Let ζ belong to $\partial_P(f_1+f_2)(x)$. Then, for every $\varepsilon > 0$, there exist $x_1, x_2 \in B(x,\varepsilon)$ with $|f_i(x_i) - f_i(x)| < \varepsilon$ $(i = 1,2)$ such that*

$$\zeta \in \partial_P f_1(x_1) + \partial_P f_2(x_2) + B(0,\varepsilon).$$

If at least one of the functions is Lipschitz near x, we have

$$\partial_L(f_1+f_2)(x) \subset \partial_L f_1(x) + \partial_L f_2(x).$$

Proof. To lighten the notation, we take $x = 0$.

A. We treat first the case $\zeta = 0$, assuming additionally that the function $f_1 + f_2$ attains a local minimum at 0. We recognize the following as a special case of the inclusion in the theorem statement:

Lemma. *For any $\varepsilon > 0$, there exist $x_1, x_2 \in B(0,\varepsilon)$ such that*

$$|f_i(x_i) - f_i(0)| < \varepsilon \ (i = 1,2) \quad \textit{and} \quad 0 \in \partial_P f_1(x_1) + \partial_P f_2(x_2) + B(0,\varepsilon).$$

Proof. Fix $\varepsilon > 0$, and let $\delta > 0$ satisfy $\delta + \delta^2 < \varepsilon$ as well as

$$|u| \leqslant \delta + \delta^2 \implies f_i(u) > f_i(0) - \varepsilon \ (i = 1,2),$$

and be such that $f_1 + f_2$ is minimized over $B(0,\delta)$ at 0. For purposes of the proof, we introduce

$$Y = \{(v,v) : v \in \mathbb{R}^n, |v| \leqslant \delta\}, \ f(x,y) = f_1(x) + f_2(y).$$

Then, in view of the local minimum, we have $\min_Y f - f(0,0) = 0$. We proceed to apply the mean value inequality (more precisely, the version given by Cor. 11.3) on $\mathbb{R}^n \times \mathbb{R}^n$, with $\varepsilon = \delta^2$, $x = (0,0)$. There results the existence of points x_1, x_2 in $B(0,\delta+\delta^2) \subset B(0,\varepsilon)$ and

$$(\zeta_1, \zeta_2) \in \partial_P f(x_1,x_2) = \partial_P f_1(x_1) \times \partial_P f_2(x_2)$$

such that

$$f_1(x_1) + f_2(x_2) \leqslant f_1(0) + f_2(0), \ \langle (\zeta_1,\zeta_2),(v,v)\rangle \geqslant -\delta^2 \ \forall (v,v) \in Y.$$

The second condition implies $|\zeta_1 + \zeta_2| \leqslant \delta < \varepsilon$. The first, given how δ was chosen, yields

$$\begin{aligned} -\varepsilon &< f_1(x_1) - f_1(0) = f_1(x_1) + f_2(x_2) - f_2(x_2) - f_1(0) \\ &\leqslant f_1(0) + f_2(0) - f_2(x_2) - f_1(0) = f_2(0) - f_2(x_2) < \varepsilon. \end{aligned}$$

Thus $|f_i(x_i) - f_i(0)| < \varepsilon \ (i = 1,2)$. The lemma is proved. □

B. We now treat the inclusion in the general case. It follows from the definition of proximal subgradient that, for a certain $\sigma \geqslant 0$, the function

$$x \mapsto f_1(x) + f_2(x) + \sigma |x|^2 - \langle \zeta, x \rangle$$

attains a local minimum at 0. We reduce the situation to that of the lemma, replacing $f_1(x)$ by

$$\widetilde{f}_1(x) = f_1(x) + \sigma |x|^2 - \langle \zeta, x \rangle.$$

Hence $\widetilde{f}_1 + f_2$ attains a local minimum at 0. Given $\varepsilon > 0$, we may now apply the lemma above for any $\varepsilon' < \varepsilon$. By taking ε' sufficiently small, and with the help of Prop. 7.31, we obtain the required conclusion; we omit the details.

C. Suppose now that f_1 is Lipschitz near 0, and let $\zeta \in \partial_L(f_1 + f_2)(0)$. Then, by the way ∂_L is defined, there is a sequence x_j converging to 0, with $(f_1 + f_2)(x_j)$ converging to $(f_1 + f_2)(0)$, and corresponding points $\zeta_j \in \partial_P(f_1 + f_2)(x_j)$ which converge to ζ. Note that $f_1(x_j) \to f_1(0)$, since f_1 is continuous; therefore, we also have $f_2(x_j) \to f_2(0)$.

Let ε_j be a positive sequence decreasing to 0. We apply the inclusion proved above to write

$$\zeta_j = \zeta_1^j + \zeta_2^j + u_j\,, \text{ where } \zeta_1^j \in \partial_P f_1\big(x_1^j\big)\,,\ \zeta_2^j \in \partial_P f_2\big(x_2^j\big)\,,$$

for certain points $u_j, x_1^j,\ x_2^j$ satisfying

$$u_j \in \varepsilon_j B\,,\ x_i^j \in B(x_j,\varepsilon_j)\,,\ \ |f_i(x_i^j) - f_i(x_j)| < \varepsilon_j\ (i = 1,2)\,.$$

Since f_1 is Lipschitz near 0, the sequence ζ_1^j is bounded; we may suppose $\zeta_1^j \to \zeta_1$ by taking a subsequence (without relabeling). Necessarily, then, ζ_2^j also converges to a limit, denoted by ζ_2. We have $\zeta = \zeta_1 + \zeta_2$. We find that, by definition, ζ_i belongs to $\partial_L f_i(0)$ $(i = 1,2)$, which completes the proof. □

11.17 Exercise. Let $F : \mathbb{R}^n \to \mathbb{R}^m$ be Lipschitz near x. Then

$$x_i \to x,\ \ \lambda_i \to \lambda,\ \ \zeta_i \to \zeta,\ \ \zeta_i \in \partial_P\langle \lambda_i, F\rangle(x_i)\ \ \forall i \implies \zeta \in \partial_L\langle \lambda, F\rangle(x). \ \square$$

Dini derivates. Differentiable functions, as well as convex functions, admit directional derivatives, as the reader knows. This is not the case for merely lower semicontinuous (or even Lipschitz) functions, however. A useful tool in such a context is the following generalized directional derivative, whose definition (somewhat extended here) can be traced back to Ulisse Dini's influential nineteenth century book on analysis.

11.18 Definition. *Let $x \in \operatorname{dom} f$, where $f : \mathbb{R}^n \to \mathbb{R}_\infty$. The (lower)* **Dini derivate** *in the direction v is given by*

$$df(x;v) = \liminf_{\substack{w \to v \\ t \downarrow 0}} \frac{f(x+tw) - f(x)}{t}\,.$$

11.19 Exercise. Let $f : \mathbb{R}^n \to \mathbb{R}$ be given.

(a) Let f be Lipschitz near x. Prove that $df(x;v)$ is given by a simpler expression:

$$df(x;v) = \liminf_{t \downarrow 0} \frac{f(x+tv) - f(x)}{t}\,.$$

(b) Let f be differentiable at x. Prove that $df(x;v) = \langle f'(x), v\rangle\ \forall\, v \in \mathbb{R}^n$. □

Subbotin's theorem. It turns out that if, from a given point, a function has a certain minimal rate of increase in all directions taken from a convex set (as measured by Dini derivates), then there is a nearby proximal subgradient that reflects that increase uniformly. This is an important fact in proximal analysis, one whose precise statement follows.

11.20 Theorem. (Subbotin) *Let $f : \mathbb{R}^n \to \mathbb{R}_\infty$ be lsc and let $x \in \operatorname{dom} f$. Suppose that we have*

$$df(x;w) > \rho \quad \forall w \in W,$$

where W is a compact convex subset of $\mathbb{R}^n$ and $\rho \in \mathbb{R}$. Then, for every $\varepsilon > 0$, there exists $x_* \in B(x,\varepsilon)$ with $|f(x_*) - f(x)| < \varepsilon$ such that, for some $\zeta \in \partial_P f(x_*)$, one has

$$\langle \zeta, w \rangle > \rho \quad \forall w \in W.$$

Proof. We claim that, for all sufficiently small $t > 0$, we have

$$f\big(x + tw + t^2 u\big) - f(x) > \rho t + t^2 \quad \forall w \in W, \quad \forall u \in B. \tag{1}$$

Were this not so, there would be sequences $t_i \downarrow 0$, $w_i \in W$, $u_i \in B$ such that

$$\frac{f\big(x + t_i w_i + t_i^2 u_i\big) - f(x)}{t_i} \leqslant \rho + t_i.$$

Invoking the compactness of W, there is a subsequence (we do not relabel) of w_i converging to a point $w \in W$. Then $w_i + t_i u_i$ converges to w as well, and we deduce $df(x;w) \leqslant \rho$, contradicting the hypothesis and establishing the claim.

Now fix $\varepsilon > 0$, and choose $t > 0$ so that (1) holds, as well as

$$tW + t^2 B \subset \varepsilon B, \ |\rho t + t^2| < \varepsilon, \ f(y) > f(x) - \varepsilon \ \forall y \in x + tW + t^2 B.$$

We proceed to apply the mean value inequality (Theorem 11.2) with data

$$Y = x + tW, \ r = \rho t + t^2, \ \varepsilon = t^2/2.$$

The point x_* that results satisfies

$$x_* \in [x, x + tW] + (t^2/2)B \subset x + tW + t^2 B \subset x + \varepsilon B,$$

so that $f(x_*) > f(x) - \varepsilon$. We also have

$$f(x_*) \leqslant f(x) + |\rho t + t^2| < f(x) + \varepsilon,$$

whence $|f(x_*) - f(x)| < \varepsilon$. Finally, the element $\zeta \in \partial_P f(x_*)$ satisfies

$$\rho t + t^2 \leqslant \langle \zeta, tw \rangle + t^2/2 \ \ \forall w \in W.$$

We find that all the requirements of the theorem statement are met. □

11.21 Exercise. Show that Subbotin's theorem is false in the absence of the convexity hypothesis on W. □

11.22 Corollary. *If f is differentiable at x, then, for every $\varepsilon > 0$, there exists a point $x_* \in B(x,\varepsilon)$ and $\zeta \in \partial_P f(x_*)$ such that $|\zeta - f'(x)| \leqslant \varepsilon$.*

Proof. The hypotheses of Subbotin's theorem are satisfied at x by the data

$$\widetilde{f}(y) = f(y) - \langle f'(x), y \rangle, \; W = B, \; \rho = -\varepsilon,$$

since $d\widetilde{f}(x;v) = 0 \;\forall v$. Applying the theorem yields a point $x_* \in B(x,\varepsilon)$, together with

$$\psi \in \partial_P \widetilde{f}(x_*) = \partial_P f(x_*) - f'(x)$$

(by Prop. 7.31) such that $\langle \psi, u \rangle > -\varepsilon \;\forall u \in B$. Then $\psi = \zeta - f'(x)$ for some ζ in $\partial_P f(x_*)$. It follows that $|\zeta - f'(x)| < \varepsilon$. □

We now see that the limiting subdifferential generates the generalized gradient, in the following sense:

11.23 Proposition. *If $f : \mathbb{R}^n \to \mathbb{R}$ is Lipschitz near x, then $\partial_C f(x) = \operatorname{co} \partial_L f(x)$. If, in addition, f is regular at x (in particular, if f is convex or continuously differentiable), then $\partial_C f(x) = \partial_L f(x)$.*

Proof. We know that $\partial_C f(x) \supset \partial_L f(x)$ by Prop. 11.12. To prove the opposite inclusion, it suffices, in light of the gradient formula for $\partial_C f(x)$ (Theorem 10.27), to show that, for some $\delta > 0$,

$$y \in B(x,\delta), \; f'(y) \text{ exists}, \; \varepsilon > 0 \implies \exists\, z \in B(y,\varepsilon), \; \zeta \in \partial_P f(z) \text{ such that } |\zeta - f'(y)| \leqslant \varepsilon.$$

But we know this, by Cor. 11.22.

Now suppose that f is regular at x, and let $\zeta \in \partial_C f(x)$. We wish to prove that $\zeta \in \partial_L f(x)$. Define $\widetilde{f}(y) = f(y) - \langle \zeta, y \rangle$. Then, for all $v \in W := B$, we have

$$d\widetilde{f}(x;v) = \widetilde{f}'(x;v) = f'(x;v) - \langle \zeta, v \rangle = f^\circ(x;v) - \langle \zeta, v \rangle \geqslant 0,$$

since $\zeta \in \partial_C f(x)$. We invoke Theorem 11.20 to deduce the existence, for any $\varepsilon > 0$, of $z \in B(x,\varepsilon)$ and $\theta \in \partial_P \widetilde{f}(z) = \partial_P f(z) - \zeta$ such that

$$\langle \theta, v \rangle \geqslant -\varepsilon \;\forall v \in B.$$

Writing $\theta = \xi - \zeta$ for some element $\xi \in \partial_P f(z)$, we obtain

$$\langle \xi - \zeta, v \rangle \geqslant -\varepsilon \;\forall v \in B,$$

whence $|\xi - \zeta| \leqslant \varepsilon$. Letting ε decrease to 0, we discover $\zeta \in \partial_L f(x)$. □

11.24 Example. (Exercise) Although ∂_P may be empty at some points and can be computationally difficult to use (compared to ∂_C), it offers certain theoretical advantages, for example in yielding the most refined characterization of Lipschitz behavior (see Cor. 11.7): the smaller the subdifferential in such a characterization, and the more often it may be empty, the fewer points there are to check.

The "importance of being empty" will now be illustrated with an example involving generalized solutions of a simple ordinary differential equation. (This theme will be extended and developed at greater length in Chapter 19, in connection with the Hamilton-Jacobi equation.) We consider the boundary-value problem

$$|\varphi'(t)|^2 = 1,\ 0 < t < 1,\ \varphi(0) = \varphi(1) = 0, \tag{$*$}$$

where $\varphi : [0,1] \to \mathbb{R}$. There is no doubt that the most desirable type of solution would be a function φ continuous on $[0,1]$ and differentiable in $(0,1)$ for which the equation $|\varphi'(t)|^2 = 1$ holds at every $t \in (0,1)$. This is a *classical solution*.

(a) Show that no classical solution of $(*)$ exists.

This is the trouble with classical solutions: we like them, but they may not be there for us. An alternative (and familiar) solution concept is the following: a Lipschitz function φ on $[0,1]$ is said to be an *almost everywhere solution* of $(*)$ if φ satisfies the boundary conditions, and if $|\varphi'(t)|^2 = 1$ a.e. (recall that $\varphi'(t)$ exists for almost every $t \in (0,1)$). With this solution concept, we gain existence. In fact, too much existence:

(b) Show that there exist an infinite number of almost everywhere solutions of $(*)$.

We now define a *proximal solution* of $(*)$: a continuous function $\varphi : [0,1] \to \mathbb{R}$ such that $\varphi(0) = \varphi(1) = 0$ and

$$t \in (0,1),\ \zeta \in \partial_P \varphi(t) \implies |\zeta|^2 = 1.$$

We proceed to prove that there is exactly one proximal solution, the function

$$\varphi_*(t) = \begin{cases} t & \text{if } t \in [0,1/2] \\ 1-t & \text{if } t \in (1/2,1]. \end{cases}$$

(c) Show that φ_* is a proximal solution of $(*)$.

Now let φ be any proximal solution of $(*)$; we show in the following steps that $\varphi = \varphi_*$.

(d) Prove that $\varphi(t) - t$ is decreasing and $\varphi(t) + t$ is increasing. Deduce that $\varphi \leqslant \varphi_*$.

(e) Prove that
$$\min_{t \in [0,1]} \varphi(t) \leqslant \varphi(1/2) - 1/2.$$
[Hint: apply the mean value inequality with $Y = [\delta, 1-\delta]$ and $x = 1/2$.]

(f) Deduce that $\varphi(1/2) = 1/2$, and conclude. [Hint: if $\varphi(1/2) < 1/2$, then φ attains a minimum in $(0,1)$, by the preceding.]

(g) Why is $-\varphi_*$ not a proximal solution? (The answer to this question illustrates the importance of ∂_P being empty.)

□

The analysis above highlighted the use of *monotonicity*, a subject that plays an important role in differential equations and control, and one for which proximal analysis is ideally suited. We shall examine the topic of monotone trajectories in the next chapter.

11.2 Proximal geometry

We now develop the geometric aspect of proximal analysis. In this section, S is always taken to be a nonempty closed subset of $\mathbb{R}^n$. The starting point is the following concept.

11.25 Definition. *Let $x \in S$. A vector $\zeta \in \mathbb{R}^n$ is a* **proximal normal** *to the set S at the point x if and only if there exists $\sigma = \sigma(x, \zeta) \geqslant 0$ such that*

$$\langle \zeta, u - x \rangle \leqslant \sigma |u - x|^2 \quad \forall u \in S. \tag{1}$$

The set $N_S^P(x)$ of all such ζ defines the proximal normal cone to S at x.

The inequality that appears above is referred to as the *proximal normal inequality*. It is evident that $N_S^P(x)$ is a convex cone containing 0.

11.26 Proposition. *We have $N_S^P(x) \subset N_S(x)$.*

Proof. Let ζ satisfy the proximal normal inequality. Since $N_S(x)$ is defined as $T_S(x)^\triangle$, we must show that $\langle \zeta, v \rangle \leqslant 0$, where v is any element of $T_S(x)$. Now we have (by definition) $v = \lim_i (x_i - x)/t_i$, where x_i is a sequence in S converging to x. For each i, we have $\langle \zeta, x_i - x \rangle \leqslant \sigma |x_i - x|^2$. Dividing by t_i and passing to the limit, we deduce $\langle \zeta, v \rangle \leqslant 0$. □

The following result confirms that, despite the global nature of the proximal normal inequality, proximal normals are a local construct.

11.27 Proposition. *Suppose that, for some $\sigma \geqslant 0$, for some positive δ, we have*

$$\langle \zeta, u - x \rangle \leqslant \sigma |u - x|^2 \quad \forall u \in S \cap B(x, \delta).$$

Then $\zeta \in N_S^P(x)$.

Proof. If the conclusion fails, then for each integer i there is a point $u_i \in S$ such that $\langle \zeta, u_i - x \rangle > i |u_i - x|^2$. It follows that $u_i \to x$. But then, when i is sufficiently large so that $|u_i - x| < \delta$ and $i > \sigma$, a contradiction ensues. □

The reader will no doubt recall that the indicator I_S of S is the function that equals 0 on S and $+\infty$ elsewhere.

11.28 Proposition. *Let $x \in S$. Then $\zeta \in N_S^P(x) \iff \zeta \in \partial_P I_S(x)$, and*

$$N_S^P(x) = \{ \lambda\zeta : \lambda \geqslant 0,\ \zeta \in \partial_P d_S(x) \}.$$

Proof. The equivalence is essentially Prop. 11.27; we turn now to the stated equality. Let $\zeta \in N_S^P(x)$. Then, for a certain $\sigma \geqslant 0$, the function

$$y \mapsto \varphi(y) := -\langle \zeta, y \rangle + \sigma |y-x|^2$$

attains a minimum relative to S at $y = x$. (This is an evident interpretation of the proximal normal inequality.) Fix any $\varepsilon > 0$. Then, on a sufficiently small neighborhood of x, the function φ is Lipschitz of rank $|\zeta| + \varepsilon$. It follows from Prop. 10.31 that x is a local minimum of the function $\varphi(y) + (|\zeta| + \varepsilon) d_S(y)$, whence

$$0 \in \partial_P \{ -\langle \zeta, y \rangle + \sigma |y-x|^2 + (|\zeta| + \varepsilon) d_S(y) \}(x) = -\zeta + (|\zeta| + \varepsilon) \partial_P d_S(x),$$

by Prop. 7.31. Thus, $\zeta/(|\zeta| + \varepsilon) \in \partial_P d_S(x)$ for every $\varepsilon > 0$. This shows that the left side of the desired set equality is contained in the right. The opposite inclusion follows easily from the definition of $\zeta \in \partial_P d_S(x)$, which immediately implies the proximal normal inequality. □

Projection generates proximal normals. The next result shows that proximal normals at x correspond to "closest point" directions emanating (outwards) from the point x, and that they are generated by projection onto the set. (Need we remind the reader that $\text{proj}_S(y)$ is the set of points $s \in S$ such that $|y-s| = d_S(y)$?)

11.29 Proposition. *A nonzero vector ζ satisfies the proximal normal inequality (1) if and only if $x \in \text{proj}_S(y)$, where $y := x + \zeta/(2\sigma)$. More generally, ζ lies in $N_S^P(x)$ if and only if there is a point $z \notin S$ for which $x \in \text{proj}_S(z)$ and such that $\zeta = t(y-x)$ for some $t > 0$.*

Proof. We have

$$\begin{aligned} x \in \text{proj}_S(x + \zeta/(2\sigma)) &\iff |\zeta/(2\sigma)| \leqslant |x + \zeta/(2\sigma) - y| \ \ \forall y \in S \\ &\iff |\zeta/(2\sigma)|^2 \leqslant |x + \zeta/(2\sigma) - y|^2 \ \ \forall y \in S \\ &\iff 0 \leqslant |x-y|^2 + \langle \zeta/\sigma, x-y \rangle \ \ \forall y \in S \end{aligned}$$

(by expanding the squared norm)

$$\iff \langle \zeta, y-x \rangle \leqslant \sigma |x-y|^2 \ \ \forall y \in S.$$

This proves the first assertion. The characterization of $N_S^P(x)$, which follows from it, is left as an exercise. □

The situation is illustrated in Fig. 11.2. Note that for all $\varepsilon > 0$ sufficiently small, the point $y - \varepsilon\zeta$ lies outside S and has *unique* closest point x in S.

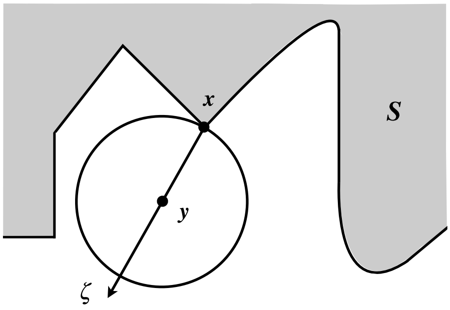


Fig. 11.2 A proximal normal direction ζ to S at x

The horizontal approximation theorem. We proceed to study the important case of proximal normals to the epigraph of a function $f : \mathbb{R}^n \to \mathbb{R}_\infty$. In the remainder of this section, f denotes a lower semicontinuous function; recall that its epigraph $\operatorname{epi} f$ is then a closed set.

11.30 Exercise. Let $x \in \operatorname{dom} f$, and suppose that $(\zeta, -\lambda) \in N^P_{\operatorname{epi} f}(x,r)$. Prove that $\lambda \geqslant 0$. If $\lambda > 0$, show that $r = f(x)$ necessarily. □

11.31 Theorem. (Rockafellar) *Let $x \in \operatorname{dom} f$, and let $\zeta \neq 0$ satisfy*

$$(\zeta, 0) \in N^P_{\operatorname{epi} f}(x,r).$$

Then, for every $\varepsilon > 0$, there exist

$$x_\varepsilon \in B(x,\varepsilon),\ \lambda_\varepsilon \in (0,\varepsilon),\ \zeta_\varepsilon \in B(\zeta,\varepsilon)$$

such that $|f(x_\varepsilon) - f(x)| < \varepsilon$ and $(\zeta_\varepsilon, -\lambda_\varepsilon) \in N^P_{\operatorname{epi} f}(x_\varepsilon, f(x_\varepsilon))$.

Proof.

A. We may assume, without loss of generality, that $|\zeta| = 1$. There exists $\delta > 0$ such that the point $(x+\delta\zeta, r)$ has *unique* closest point (x,r) in $\operatorname{epi} f$ (see the remark following Prop. 11.29):

$$|(u, f(u)+\varepsilon) - (x+\delta\zeta, r)| \geqslant |(x,r) - (x+\delta\zeta, r)| = \delta \ \ \forall u \in \operatorname{dom} f,\ \varepsilon \geqslant 0.$$

Substituting $\varepsilon + (r - f(x))$ for ε in this inequality, we obtain

$$|(u, f(u)+\varepsilon) - (x+\delta\zeta, f(x))| \geqslant \delta \ \ \forall u \in \operatorname{dom} f,\ \varepsilon \geqslant 0.$$

Thus, $(\zeta,0) \in N^P_{\operatorname{epi} f}(x, f(x))$, by Prop. 11.29. Note also that the conclusion of the theorem is unrelated to the value of r. The moral of this argument: it suffices to prove the theorem in the case $r = f(x)$, which we proceed to do.

B. Let $t > 0$. Then for any $(u,\lambda) \in \operatorname{epi} f$, we have, by the uniqueness,

$$\begin{aligned} |(x+\delta\zeta, f(x)-t)-(u,\lambda)| &= |(x+\delta\zeta, f(x))-(u,\lambda+t)| \\ &> |(x+\delta\zeta, f(x))-(x,f(x))|. \end{aligned}$$

Letting d be the distance function for $\operatorname{epi} f$, the conclusion is that

$$d(x+\delta\zeta, f(x)-t) > d(x+\delta\zeta, f(x)) \;\; \forall\, t > 0.$$

It follows that there exists a sequence $(x_i, t_i) \to (x+\delta\zeta, 0)$ with $t_i > 0$ such that $\nabla d(x_i, f(x)-t_i)$ exists and has strictly negative second component. We know (see Exer. 10.40) that

$$\nabla d(x_i, f(x)-t_i) = (x_i - y_i, f(x)-t_i-r_i)/d(x_i, f(x)-t_i),$$

where (y_i, r_i) is a closest point in $\operatorname{epi} f$ to $(x_i, f(x)-t_i)$; we have $r_i = f(y_i)$ necessarily (since $r_i > f(x)-t_i$). Thus

$$(x_i - y_i, f(x)-t_i-f(y_i))/d(x_i, f(x)-t_i) \in N^P_{\operatorname{epi} f}\big(y_i, f(y_i)\big),$$

and $(y_i, f(y_i)) \to (x, f(x))$ necessarily (by uniqueness of the closest point). But

$$\big(x_i - y_i, f(x)-t_i-f(y_i)\big)/d\big(x_i, f(x)-t_i\big) \to (\delta\zeta, 0)/\delta = (\zeta, 0),$$

and the result follows. □

The following reflects the familiar fact from classical calculus that the vector $(\nabla f(x), -1)$ is a downward-pointing normal vector to the graph of f.

11.32 Theorem. *Let* $x \in \operatorname{dom} f$. *Then*

$$\zeta \in \partial_P f(x) \iff (\zeta, -1) \in N^P_{\operatorname{epi} f}\big(x, f(x)\big).$$

Proof. Suppose first that $\zeta \in \partial_P f(x)$. Then, by definition, there exist $\sigma \geqslant 0$ and a neighborhood V of x such that

$$f(y) - f(x) + \sigma|y-x|^2 \geqslant \langle \zeta, y-x \rangle \;\; \forall\, y \in V.$$

Rewriting yields

$$\langle (\zeta,-1), (y-x, f(y)-f(x)) \rangle \leqslant \sigma|y-x|^2 \;\; \forall\, y \in V,$$

which in turn implies

$$\langle (\zeta,-1), [(y,\alpha)-(x,f(x))] \rangle \leqslant \sigma |(y,\alpha)-(x,f(x))|^2 \qquad (2)$$

for all points $(y,\alpha) \in \operatorname{epi} f$ in a neighborhood of $(x,f(x))$. It follows that $(\zeta,-1)$ belongs to $N^P_{\operatorname{epi} f}(x,f(x))$.

Let us now turn to the converse: suppose that $(\zeta,-1) \in N^P_{\operatorname{epi} f}(x,f(x))$. Then, by definition, there exists $\sigma \geqslant 0$ such that (2) holds for all $(y,\alpha) \in \operatorname{epi} f$. Now fix M greater than $\sigma(1+|\zeta|^2)$. We claim that, for all y in a neighborhood of x, we have

$$f(y)-f(x)+M|y-x|^2 \geqslant \langle \zeta, y-x \rangle.$$

If this is not the case, there is a sequence $y_i \to x$ for which the inequality fails:

$$f(y_i)-f(x)+M|y_i-x|^2 < \langle \zeta, y_i-x \rangle.$$

Note that $y_i \neq x$ necessarily. Now set

$$\alpha_i = f(x)+\langle \zeta, y_i-x \rangle - M|y_i-x|^2.$$

By the preceding inequality, we have $(y_i,\alpha_i) \in \operatorname{epi} f$. Substituting in (2), we derive

$$(M-\sigma)|y_i-x|^2 \leqslant \sigma \big| \langle \zeta, y_i-x \rangle - M|y_i-x|^2 \big|^2.$$

Dividing across by $|y_i-x|^2$ and letting $i \to \infty$, we deduce $M-\sigma \leqslant \sigma|\zeta|^2$, which contradicts the way M was chosen. □

Given the geometrical meaning of proximal normals (closest points) and proximal subgradients (locally supporting parabolas), we may interpret the theorem above as saying that when the underlying set is an epigraph, it is equivalent to have a contact *sphere* and a contact *parabola* at a given point $(x,f(x))$, provided that the normal direction is non horizontal.

Limiting normals. We define the *limiting normal cone* $N^L_S(x)$ by means of a closure operation applied to N^P_S:

$$N^L_S(x) = \big\{ \zeta = \lim_{i\to\infty} \zeta_i : \zeta_i \in N^P_S(x_i),\ x_i \to x,\ x_i \in S \big\}.$$

(Strictly speaking, it is superfluous to say that the points x_i lie in S, since the normal cone $N^P_S(x_i)$ is not defined otherwise.) The cone $N^P_S(x)$ is always convex (by definition), but may not be closed; the cone $N^L_S(x)$ is always closed (by construction), but may not be convex.

11.33 Exercise. Consider the set S in $\mathbb{R}^2$ of Exer. 10.41. Prove that

$$N^L_S(0,0) = \big\{ (\delta,0) : \delta \geqslant 0 \big\} \cup \big\{ (0,\lambda) : \lambda \geqslant 0 \big\}.$$

Observe that $N_S(0,0)$, $N^L_S(0,0)$, and $N^C_S(0,0)$ are all different. □

11.34 Proposition. *Let $x \in S$. Then*

$$\zeta \in N_S^L(x) \iff \zeta \in |\zeta| \partial_L d_S(x),$$

and

$$N_S^L(x) = \{\lambda \zeta : \lambda \geqslant 0, \zeta \in \partial_L d_S(x)\} = \partial_L I_S(x).$$

Proof. Let $0 \neq \zeta \in N_S^L(x)$. Then there is a sequence ζ_i converging to ζ and a sequence x_i in S converging to x such that $\zeta_i \in N_S^P(x_i)$. The proof of Prop. 11.28 showed that $\zeta_i/(|\zeta_i|+\varepsilon) \in \partial_P d_S(x_i)$, for any $\varepsilon > 0$. We deduce that $\zeta/|\zeta|$ belongs to $\partial_L d_S(x)$.

For the converse, we may restrict attention to the case $|\zeta| = 1$. There exists a sequence x_i converging to x and a sequence $\zeta_i \in \partial_P d_S(x_i)$ converging to ζ. If $x_i \in S$ infinitely often, then $\zeta_i \in N_S^P(x_i)$ by Prop. 11.28, whence $\zeta \in N_S^L(x)$. In the other case, we have $x_i \notin S$ for all i sufficiently large. It follows from Prop. 7.39 that $\zeta_i = (x_i - s_i)/d_S(x_i)$, where $\text{proj}_S(x_i) = s_i$. Then, by Prop. 11.29, we have $\zeta_i \in N_S^P(s_i)$. Since $s_i \to x$, we deduce $\zeta \in N_S^L(x)$.

The remaining assertions of the proposition are left as a simple exercise. □

11.35 Corollary. *Let x be a local minimum of the function $f(u)$ subject to the constraint $u \in S$, where f is Lipschitz near x. Then $0 \in \partial_L f(x) + N_S^L(x)$.*

Proof. The function $f + I_S$ has a local minimum at x, whence

$$0 \in \partial_P(f + I_S)(x) \subset \partial_L(f + I_S)(x).$$

We invoke the sum rule (Theorem 11.16) and the theorem above to conclude. □

11.36 Theorem. *Let $x \in S$. Then*

$$N_S(x) \subset N_S^L(x) \subset N_S^C(x) = \overline{\text{co}}\, N_S^L(x),$$

with equality if and only if S is regular at x.

Proof. Recall (Theorem 10.34) that

$$N_S^C(x) = \overline{\text{co}}\{\lambda \zeta : \lambda \geqslant 0, \zeta \in \partial_C d_S(x)\}.$$

By Prop. 11.23, we also have $\partial_C d_S(x) = \text{co}\, \partial_L d_S(x)$. Then, in view of Prop. 11.34, we may write

$$\begin{aligned} \overline{\text{co}}\, N_S^L(x) &= \overline{\text{co}}\Big\{ \bigcup\nolimits_{\lambda \geqslant 0} \lambda\, \partial_L d_S(x) \Big\} = \overline{\text{co}}\Big\{ \bigcup\nolimits_{\lambda \geqslant 0} \text{co}\big[\lambda\, \partial_L d_S(x)\big] \Big\} \\ &= \overline{\text{co}}\Big\{ \bigcup\nolimits_{\lambda \geqslant 0} \lambda\, \partial_C d_S(x) \Big\} = N_S^C(x). \end{aligned}$$

In order to complete the proof of the theorem, (that is, to prove the first inclusion), we need only show that any element ζ of $N_S(x)$ belongs to $N_S^L(x)+B(0,\varepsilon)$, for an arbitrary $\varepsilon > 0$, which we proceed to do.

The definition of $N_S(x)$ implies (see Exer. 8.6) the existence of a neighborhood V of x such that $\langle \zeta, u-x \rangle \leqslant \varepsilon |u-x| \;\; \forall u \in S \cap V$. Then x is a local minimum for the function $\langle -\zeta, u-x \rangle + \varepsilon |u-x|$ relative to $u \in S$. Calling upon Cor. 11.35, we may then write

$$0 \in \partial_L \big\{ \langle -\zeta, u \rangle + \varepsilon |u-x| \big\}(x) + N_S^L(x) \subset -\zeta + B(0,\varepsilon) + N_S^L(x),$$

the required conclusion.

Finally, regularity of S at x is characterized by equality between $N_S^C(x)$ and $N_S(x)$, whence the last assertion of the theorem. □

It follows that all the normal cones that we currently dispose of coincide when the underlying set is convex or sufficiently smooth (for example, a C^2 manifold). In general, they may all be different, however.

We remark that, in contrast to the nonsmooth geometry associated with generalized gradients, tangency does not enter into proximal geometry: the latter is based entirely on normality. We further remark that, even though we have taken the set S to be closed above, it would suffice (for purposes of studying normal cones) that it be locally closed near the point x of interest. This is because proximal normals are entirely determined by the *local* structure of the underlying set.

11.3 A proximal multiplier rule

It is a truth universally acknowledged that a subdifferential in possession of a good calculus, must be in want of a multiplier rule. We examine now the proximal version of this adage, in the context of the following optimization problem:

$$\text{Minimize } f(x) \text{ subject to } \varphi(x) \in \Phi, \tag{Q}$$

where the functions $f : \mathbb{R}^n \to \mathbb{R}_\infty$ and $\varphi : \mathbb{R}^n \to \mathbb{R}^k$, together with a subset Φ of $\mathbb{R}^k$, are the given data of the problem. It will always be assumed that f is lsc and Φ is closed. As pointed out in Exer. 10.48, (Q) is equivalent to the problem (P) considered earlier (while being notationally more convenient on occasion).

The following illustrates a now familiar theme: we derive a fuzzy proximal assertion, together with its limiting exact version. Note that, at the fuzzy level, the multiplier rule below always holds in *normal* form.

11.37 Theorem. *Let x_* be a local minimizer for* (Q)*, where φ is Lipschitz near x_*. Then, for any $\varepsilon > 0$, there exist*

$$x_\varepsilon \in B(x_*, \varepsilon),\ y_\varepsilon \in B\big(\varphi(x_*), \varepsilon\big) \cap \Phi,\ \nu_\varepsilon \in N_\Phi^P(y_\varepsilon)$$

with $|f(x_\varepsilon) - f(x_*)| < \varepsilon$ such that

$$0 \in \partial_P\big\{f + \langle \nu_\varepsilon, \varphi \rangle\big\}(x_\varepsilon) + B(0, \varepsilon).$$

If f is Lipschitz near x_*, there exist η equal to 0 *or* 1 *and $\nu \in N_\Phi^L(\varphi(x_*))$ such that $(\eta, \nu) \neq 0$ and*

$$0 \in \partial_L\big\{\eta f + \langle \nu, \varphi \rangle\big\}(x_*).$$

Proof.

A. Fix $r > 0$ such that φ is Lipschitz on $B(x_*, r)$, and such that x_* is optimal for (Q) relative to this ball. Let us now define $V : \mathbb{R}^k \to \mathbb{R}_\infty$ as follows:

$$V(\alpha) = \inf\big\{f(x) + |x - x_*|^2 : x \in B(x_*, r),\ \varphi(x) \in \Phi - \alpha\big\}.$$

It is clear that $V(0) = f(x_*)$, and that the infimum defining $V(\alpha)$ is attained when $V(\alpha) < \infty$; that is, when the set $\{x \in B(x_*, r) \cap \operatorname{dom} f : \varphi(x) \in \Phi - \alpha\}$ is nonempty.

Lemma. *V is lower semicontinuous.*

Proof. Let α_i be a sequence converging to α, and such that $\lim_i V(\alpha_i) \leqslant \ell \in \mathbb{R}$; we wish to prove $V(\alpha) \leqslant \ell$ (see Prop. 2.15). Let x_i be a point at which the infimum defining $V(\alpha_i)$ is attained. Then

$$x_i \in B(x_*, r),\ \varphi(x_i) \in \Phi - \alpha_i,\ V(\alpha_i) = f(x_i) + |x_i - x_*|^2.$$

By taking a subsequence, we may suppose $x_i \to \bar{x}$. Then $\bar{x}$ is admissible for the problem defining $V(\alpha)$, whence

$$V(\alpha) \leqslant f(\bar{x}) + |\bar{x} - x_*|^2 \leqslant \lim_{i\to\infty}\{f(x_i) + |x_i - x_*|^2\} = \lim_{i\to\infty} V(\alpha_i) \leqslant \ell.$$

B. According to the proximal density theorem 7.34, there is a sequence $\alpha_i \to 0$ such that

$$V(\alpha_i) < \infty,\ V(\alpha_i) \to V(0),\ \partial_P V(\alpha_i) \neq \emptyset.$$

Let $\zeta_i \in \partial_P V(\alpha_i)$, and let x_i be a point at which the minimum defining $V(\alpha_i)$ is attained. Then $f(x_i) + |x_i - x_*|^2 \to f(x_*)$; since $x_i \in B(x_*, r)\ \forall\, i$, we may suppose (by taking a subsequence) that x_i converges to a limit $\bar{x}$. It follows that $\bar{x}$ is admissible for (Q) and satisfies

$$f(\bar{x}) + |\bar{x} - x_*|^2 \leqslant f(x_*),$$

whence $\bar{x} = x_*$ (for otherwise the optimality of x_* would fail). Thus $f(x_i) \to f(x_*)$ and, of course, $\varphi(x_i) \to \varphi(x_*)$. We now fix any i sufficiently large so that x_i belongs to $B^\circ(x_*, r)$.

By the definition of proximal subgradient, there exist $\sigma_i > 0$ and a neighborhood U_i of α_i such that

$$V(\alpha) + \sigma_i |\alpha - \alpha_i|^2 - \langle \zeta_i, \alpha \rangle \geqslant V(\alpha_i) - \langle \zeta_i, \alpha \rangle \;\; \forall \alpha \in U_i. \tag{1}$$

There exists $y_i \in \Phi$ such that $\varphi(x_i) = y_i - \alpha_i$. Now let V_i be a neighborhood of x_i and W_i a neighborhood of y_i such that

$$x \in V_i, \; y \in W_i \cap \Phi \implies x \in B(x_*, r), \; y - \varphi(x) \in U_i.$$

For such a choice of x and y, and with $\alpha := y - \varphi(x)$, we have therefore

$$f(x) + |x - x_*|^2 \geqslant V(\alpha)$$

by definition of V, since $\varphi(x) = y - \alpha$ and $x \in B(x_*, r)$. We also have

$$V(x_i) = f(x_i) + |x_i - x_*|^2, \;\; \alpha \in U_i.$$

Substituting in (1), we deduce that the function

$$(x,y) \mapsto f(x) + |x - x_*|^2 + \sigma_i |y - \varphi(x) - y_i + \varphi(x_i)|^2 - \langle \zeta_i, y - \varphi(x) \rangle$$

attains a minimum over $V_i \times (W_i \cap \Phi)$ at $(x,y) = (x_i, y_i)$. The same therefore holds for the function

$$(x,y) \mapsto f(x) + |x - x_*|^2 + 2\sigma_i |y - y_i|^2 + 2\sigma_i K^2 |x - x_i|^2 - \langle \zeta_i, y - \varphi(x) \rangle,$$

where K is a Lipschitz constant for φ on $B(x_*, r)$.

Fix $y = y_i$ in the context of this minimization; we obtain from Fermat's rule

$$0 \in \partial_P \{ f + \langle \zeta_i, \varphi \rangle \} + 2(x_i - x_*). \tag{2}$$

Now fix $x = x_i$; we derive $\zeta_i \in N^P_\Phi(y_i)$. For i sufficiently large, we obtain the first part of the theorem, by taking $x_\varepsilon = x_i, \; y_\varepsilon = \varphi(x_i) + \alpha_i, \; \nu_\varepsilon = \zeta_i$.

C. We turn now to the last assertion of the theorem. If, for at least a subsequence, the ζ_i are bounded, we take a subsequence for which they converge to a limit ν, which necessarily lies in $N^L_\Phi(\varphi(x_*))$. Then we pass to the limit in (2), obtaining (see Exer. 11.17) $0 \in \partial_L \{ f + \langle \nu, \varphi \rangle \}(x_*)$. This is the desired conclusion, in the normal case.

In the remaining case, we have $\lim_i |\zeta_i| = \infty$. We then take a subsequence such that $\zeta_i / |\zeta_i| \to \nu$, and we divide across in (2) by $|\zeta_i|$ before passing to the limit.

We discover $0 \in \partial_L\{\langle \nu, \varphi \rangle\}(x_*)$, where $\nu \neq 0$; this is the abnormal version of the required conclusion. □

Normal vectors to functionally defined sets. A useful geometrical consequence of Theorem 11.37 is one bearing upon sets defined by functional constraints of the form $\varphi(x) \in \Phi$. The nondegeneracy of the constraint formulation corresponds in this setting to the *constraint qualification* postulated in the theorem below. The reader has actually met two special cases of it already: the rank condition (p. 95) or surjectivity condition for equality constraints (as in Theorem 10.45), and the positive linear independence hypothesis for inequalities (see Cor. 10.44).

11.38 Theorem. *Let E be the set defined by $\{u \in \mathbb{R}^n : \varphi(u) \in \Phi\}$, where Φ is a closed subset of $\mathbb{R}^k$, and where $\varphi : \mathbb{R}^n \to \mathbb{R}^k$ is Lipschitz near $x \in E$. We posit the following constraint qualification:*

$$0 \in \partial_L\langle \nu, \varphi \rangle(x)\,,\ \nu \in N^L_\Phi\big(\varphi(x)\big) \implies \nu = 0\,.$$

Then if $\zeta \in N^L_E(x)$, there exists $\nu \in N^L_\Phi(\varphi(x))$ such that $\zeta \in \partial_L\langle \nu, \varphi \rangle(x)$.

Proof. There exist sequences $x_i \to x,\ \zeta_i \to \zeta$ such that $\zeta_i \in N^P_E(x_i)$. By definition of N^P_E, for a certain σ_i, the point x_i is a local minimum for the function

$$x \mapsto \langle -\zeta_i\,, x\rangle + \sigma_i|x - x_i|^2$$

relative to $\varphi(x) \in \Phi$. We invoke Theorem 11.37: there exist $\eta^i = 0$ or 1, and ν_i in $N^L_\Phi(\varphi(x_i))$ such that

$$0 \in \partial_L\big\{-\eta^i\langle \zeta_i\,, x\rangle + \eta^i\sigma_i|x - x_i|^2 + \langle \nu_i\,, \varphi(x)\rangle\big\}(x_i)\,.$$

This implies $\eta^i\zeta_i \in \partial_L\langle \nu_i, \varphi \rangle(x_i)$. Dividing by $|(\eta^i, \nu_i)|$, and taking a suitable subsequence, we deduce in the limit (see Exer. 11.17) $\eta\,\zeta \in \partial_L\langle \nu_0, \varphi \rangle(x)$, where $\nu_0 \in N^L_\Phi(\varphi(x))$, $\eta \geqslant 0$, and $|(\eta, \nu_0)| = 1$. The constraint qualification implies that η is nonzero, whence

$$\zeta \in \partial_L\langle \nu_0/\eta\,, \varphi \rangle(x)\,.$$

Setting $\nu = \nu_0/\eta$, we recognize the required conclusion. □

11.39 Theorem. *Let $x \in S_1 \cap S_2$, where S_1 and S_2 are closed subsets of $\mathbb{R}^n$ which are transversal at x, in the sense that $-N^L_{S_1}(x) \cap N^L_{S_2}(x) = \{0\}$. Then*

$$N^L_{S_1 \cap S_2}(x) \subset N^L_{S_1}(x) + N^L_{S_2}(x)\,.$$

If, in addition, S_1 and S_2 are regular at x, then this holds with equality, in which case $S_1 \cap S_2$ is also regular at x, and we have $T_{S_1 \cap S_2}(x) = T_{S_1}(x) \cap T_{S_2}(x)$.

Proof. Define $\varphi(x,x) = (x,x)$ and $\Phi = S_1 \times S_2$. If $(\nu_1, \nu_2) \in N^L_{S_1}(x) \times N^L_{S_2}(x)$, then

$$0 \in \partial_L \langle (v_1, v_2), \varphi \rangle (x,x) \implies v_1 + v_2 = 0 \implies v_1 = v_2 = 0,$$

in view of transversality. This verifies the constraint qualification of Theorem 11.38. Note that the set E of that theorem is $S_1 \times S_2$. Accordingly, we have

$$\zeta \in N^L_{S_1 \cap S_2}(x) \implies \zeta \in \partial_L \langle (v_1, v_2), \varphi \rangle (x) = v_1 + v_2$$

for certain $(v_1, v_2) \in N^L_{S_1}(x) \times N^L_{S_2}(x)$, which yields the desired conclusion.

Suppose now that S_1 and S_2 are regular at x.

Lemma. *$N^L_{S_1}(x) + N^L_{S_2}(x)$ is closed and convex.*

Proof. We verify the closedness first. Let $\zeta_i + \xi_i$ be a sequence in $N^L_{S_1}(x) + N^L_{S_2}(x)$ converging to a limit ψ. If ζ_i admits a bounded subsequence, then, by taking a suitable subsequence, it follows that $\psi \in N^L_{S_1}(x) + N^L_{S_2}(x)$, since each of these cones is closed.

If, to the contrary, $|\zeta_i| \to \infty$, then we divide across by $|\zeta_i|$ and (again, for a suitable subsequence) obtain in the limit nonzero elements $\zeta \in N^L_{S_1}(x)$, $\xi \in N^L_{S_2}(x)$ whose sum is zero. This contradicts the transversality hypothesis, and proves that the sum is closed. That the sum is convex is evident, since N^L and N^C agree at regular points (Theorem 11.36), and N^C is convex.

Now we calculate (see Exer. 1.38 (c))

$$T^C_{S_1 \cap S_2}(x) \subset T_{S_1 \cap S_2}(x) \subset T_{S_1}(x) \cap T_{S_2}(x) = T^C_{S_1}(x) \cap T^C_{S_2}(x),$$

in view of the regularity. Recall that when S is regular at x, then $N_S(x)$, $N^L_S(x)$, and $N^C_S(x)$ all agree (Theorem 11.36). Taking polars above, we find

$$\begin{aligned} N^C_{S_1 \cap S_2}(x) \supset N_{S_1 \cap S_2}(x) &\supset \big[T^C_{S_1}(x) \cap T^C_{S_2}(x) \big]^{\triangle} \supset \big[T^C_{S_1}(x) \big]^{\triangle} + \big[T^C_{S_2}(x) \big]^{\triangle} \\ &= N^C_{S_1}(x) + N^C_{S_2}(x) = N_{S_1}(x) + N_{S_2}(x) = N^L_{S_1}(x) + N^L_{S_2}(x) \\ &= \overline{\text{co}} \, \{ N^L_{S_1}(x) + N^L_{S_2}(x) \} \supset \overline{\text{co}} \, N^L_{S_1 \cap S_2}(x) = N^C_{S_1 \cap S_2}(x), \end{aligned}$$

where we have used both the lemma and the inclusion established in the first part of the proof. Thus, equality holds throughout, which yields the remaining assertions of the theorem. □

11.40 Exercise. Let E be defined as $\{x \in S : g(x) \leqslant 0,\ h(x) = 0\}$, where g and h are continuously differentiable functions with values in $\mathbb{R}^m$ and $\mathbb{R}^k$ respectively, and S is a closed subset of $\mathbb{R}^n$. Suppose that the following constraint qualification holds at a point $x \in E$:

$$0 \in D_x \{ \langle \gamma, g \rangle + \langle \lambda, h \rangle \}(x) + N^L_S(x),\ \gamma \geqslant 0,\ \langle \gamma, g(x) \rangle = 0 \implies \gamma = 0,\ \lambda = 0.$$

Prove, with the help of Theorem 11.38, that if $\zeta \in N_E^L(x)$, then there exist $\gamma \in \mathbb{R}^m_+$, $\lambda \in \mathbb{R}^k$, with $\langle \gamma, g(x) \rangle = 0$, such that

$$\zeta \in D_x\{\langle \gamma, g \rangle + \langle \lambda, h \rangle\}(x) + N_S^L(x). \qquad \square$$

The proximal chain rule. Unexpectedly, perhaps, it turns out that the multiplier rule admits the chain rule as an immediate consequence.

We consider a composite function f of the form $f = g \circ h$, where $h : \mathbb{R}^n \to \mathbb{R}^k$, and where $g : \mathbb{R}^k \to \mathbb{R}_\infty$ is lower semicontinuous. We are given a point x such that $h(x) \in \operatorname{dom} g$, and we suppose that h is Lipschitz near x.

11.41 Theorem. *Let $\zeta \in \partial_P f(x)$. Then, for any $\varepsilon > 0$, there exist*

$$y_\varepsilon \in B\big(h(x), \varepsilon\big),\ \theta_\varepsilon \in \partial_P g(y_\varepsilon),\ \text{and } x_\varepsilon \in B(x, \varepsilon)$$

with $|g(y_\varepsilon) - g(h(x))| < \varepsilon$ such that $\zeta \in \partial_P \langle \theta_\varepsilon, h \rangle(x_\varepsilon) + B(0, \varepsilon)$. If g is Lipschitz near $h(x)$, then we have

$$\zeta \in \partial_L f(x) \implies \exists\, \theta \in \partial_L g\big(h(x)\big) \text{ such that } \zeta \in \partial_L \langle \theta, h \rangle(x).$$

Proof. There exists $\sigma \geqslant 0$ such that the function $u \mapsto f(u) + \sigma |u - x|^2 - \langle \zeta, u \rangle$ attains a local minimum at x. Then the function

$$(u, y) \mapsto g(y) + \sigma |u - x|^2 - \langle \zeta, u \rangle$$

attains a local minimum at $(x, h(x))$, relative to the constraint $h(u) - y = 0$. We apply Theorem 11.37 to obtain the desired conclusion. When g is Lipschitz near $h(x)$, it is straightforward to pass to the limit in order to obtain the final assertion (see Exer. 11.17). $\square$

11.4 Dini and viscosity subdifferentials

The Dini derivate $df(x; v)$ that we met in Def. 11.18 leads to a natural subdifferential construct.

11.42 Definition. *Let $f : \mathbb{R}^n \to \mathbb{R}_\infty$ and $x \in \operatorname{dom} f$. The* **Dini subdifferential** *of f at x, denoted $\partial_D f(x)$, consists of the elements $\zeta \in \mathbb{R}^n$ such that*

$$df(x; v) \geqslant \langle \zeta, v \rangle \ \ \forall\, v \in \mathbb{R}^n.$$

Each such ζ is referred to as a Dini subgradient.

This is at least the third subdifferential concept to enter the lists; the reader will soon understand our reasons for introducing it. We proceed now to a brief summary of the main features of $\partial_D f$, beginning with the following, whose proof is left as an exercise.

11.43 Proposition. *Let* $f:\mathbb{R}^n \to \mathbb{R}_\infty$ *and* $x \in \operatorname{dom} f$.

(a) ζ *belongs to* $\partial_D f(x)$ *if and only if there exists a function* $o(\cdot):\mathbb{R}_+ \to \mathbb{R}_+$ *such that* $\lim_{r\downarrow 0} o(r)/r = 0$ *and*

$$f(x+u) - f(x) - \langle\, \zeta, u\,\rangle + o(|\,u\,|) \geqslant 0 \;\; \forall\, u \in \mathbb{R}^n;$$

(b) $\partial_D f(x)$ *is closed and convex, and contains* $\partial_P f(x)$;

(c) *If* f *is differentiable at* x, *then* $\partial_D f(x) = \{f'(x)\}$.

Note that, in contrast to ∂_P, ∂_L, or ∂_C, the Dini subdifferential ∂_D always reduces to the derivative when it exists. There is no duality here between the directional and subdifferential constructs, however; that is, we cannot reconstruct $df(x;\cdot)$ from $\partial_D f(x)$, in contrast to the pair $f^\circ(x;\cdot)$ and $\partial_C f(x)$, where the former is the support function of the latter.

11.44 Proposition. *We have* $\zeta \in \partial_D f(x)$ *if and only if there exists a continuous function* $g\colon \mathbb{R}^n \to \mathbb{R}$ *which is differentiable at* x, *with* $g'(x) = \zeta$, *such that* $f-g$ *attains a local minimum at* x.

Proof. If a function g as described exists, then we have

$$d(f-g)(x;v) = df(x;v) - \langle g'(x), v\rangle \geqslant 0 \;\; \forall\, v,$$

whence $\zeta = g'(x) \in \partial_D f(x)$. We turn now to the converse: let $\zeta \in \partial_D f(x)$.

Let us define

$$o_1(t) = t \sup\big\{ o(r)/r : \; 0 < r \leqslant t \big\} \;\geqslant\; o(t),$$

where o is the function provided by Prop. 11.43. It is not hard to show that o_1 satisfies the same properties as o, but with the additional property that $t \mapsto o_1(t)/t$ is increasing on $(0,+\infty)$ (and hence continuous except at countably many points). Next, we set

$$o_2(t) \;=\; 2t \int_t^{2t} \frac{o_1(r)}{r^2}\, dr.$$

The reader may verify (using the fact that $t \mapsto o_1(t)/t$ is increasing) that

$$o_1(2t) \;\geqslant\; o_2(t) \;\geqslant\; o_1(t) \;\; \forall\, t > 0,$$

from which it follows that o_2 satisfies all the properties mentioned for o_1, in addition to being continuous. Observe that

$$f(y) \geqslant g(y) := f(x) + \langle \zeta, y - x \rangle - o_2(|y - x|) \ \forall y \in \mathbb{R}^n,$$

with equality at $y = x$ (since $o_2(0) = 0$). Thus $f - g$ attains a minimum at x. Clearly g is continuous.

There remains to prove that $g'(x)$ exists and equals ζ; that is, that the function $w(y) = o_2(|y - x|)$ satisfies $w'(x) = 0$. But this is evident, since, for any $y \neq x$, we have

$$0 \leqslant \frac{w(y) - w(x)}{|y - x|} = \frac{o_2(|y - x|)}{|y - x|},$$

and the last term converges to 0 as $y \to x$. □

The characterization in Prop. 11.44 is the one most often used in the literature of viscosity solutions, which explains why **viscosity subdifferential** is a synonym for the Dini subdifferential $\partial_D f$.

We proceed to show, using Subbotin's theorem, that even though $\partial_D f(x)$ is sometimes strictly bigger than $\partial_P f(x)$, the difference between the two is negligible, in a certain sense.

11.45 Theorem. *Let $f : \mathbb{R}^n \to \mathbb{R}_\infty$ be lsc. Suppose that $\zeta \in \partial_D f(x)$. Then, for any $\varepsilon > 0$, there exist $z \in x + \varepsilon B$ and $\xi \in \partial_P f(z)$ such that*

$$|f(x) - f(z)| < \varepsilon, \ |\zeta - \xi| < \varepsilon.$$

Proof. Set $\varphi(y) = f(y) - \langle \zeta, y \rangle$. Then $d\varphi(x; v) \geqslant 0 \ \forall v \in \mathbb{R}^n$. Thus, for any $\delta > 0$, we have

$$d\varphi(x; v) > -\delta \ \forall v \in B.$$

We apply Subbotin's theorem 11.20: there exists z and $\psi \in \partial_P \varphi(z)$ such that

$$z \in x + \delta B, \quad |\varphi(z) - \varphi(x)| < \delta, \ \langle \psi, v \rangle > -\delta \ \forall v \in B.$$

It follows that $|\psi| < \delta$ and

$$|f(z) - f(x)| \leqslant |\varphi(z) - \varphi(x)| + |\zeta||x - z| < (1 + |\zeta|)\delta.$$

We also have $\partial_P \varphi(z) = \partial_P f(z) - \zeta$, by Prop. 7.31, which implies the existence of $\xi \in \partial_P f(z)$ such that $\psi = \xi - \zeta$. Then, choosing $\delta < \varepsilon/(1 + |\zeta|)$, we find that z and ξ satisfy the required conditions. □

11.46 Corollary. *Let f be differentiable at x. Then, for any positive ε, there exists $z \in x + \varepsilon B$ admitting $\zeta \in \partial_P f(z)$ such that $|f(x) - f(z)| < \varepsilon$ and $|f'(x) - \zeta| < \varepsilon$.*

A further consequence of the theorem is that when we apply the sequential closure operation to $\partial_D f$ as we did for $\partial_P f$, the resulting limiting construct is the *same* as before:

11.47 Corollary. *Let $f : \mathbb{R}^n \to \mathbb{R}_\infty$ be lsc. Then we have*

$$\partial_L f(x) = \big\{ \zeta = \lim_{i\to\infty} \zeta_i \,:\, \zeta_i \in \partial_D f(x_i),\ x_i \to x,\ f(x_i) \to f(x) \big\}.$$

It also follows from Theorem 11.45 that $\partial_D f$ possesses the *same* fuzzy calculus as ∂_P. (One fuzziness subsumes the other, so to speak.) To give but one example, we state the following sum rule, which now follows immediately from the proximal sum rule given by Theorem 11.16.

11.48 Corollary. *Let $f_1, f_2 : \mathbb{R}^n \to \mathbb{R}_\infty$ be lsc, with $x \in \operatorname{dom} f_1 \cap \operatorname{dom} f_2$. Let ζ belong to $\partial_D(f_1 + f_2)(x)$. Then, for every $\varepsilon > 0$, there exist $x_1, x_2 \in B(x,\varepsilon)$ such that*

$$|\, f_i(x_i) - f_i(x)| < \varepsilon \ \ (i = 1,2) \ \text{ and } \ \zeta \in \partial_D f_1(x_1) + \partial_D f_2(x_2) + B(0,\varepsilon)\,.$$

A moral of the discussion is that (in finite dimensions) the development of subdifferential calculus can just as well be based on ∂_D (instead of ∂_P).

We conclude by showing that the Dini subdifferential can be used to characterize regularity at a point.

11.49 Theorem. *If $f : \mathbb{R}^n \to \mathbb{R}$ is Lipschitz near x, then*

$$\partial_D f(x) \subset \partial_L f(x) \subset \partial_C f(x)\,,$$

with equality throughout if and only if f is regular at x.

Proof. The inclusions are known facts. Suppose now that f is regular at x, and let ζ belong to $\partial_C f(x)$. Then (using Exer. 11.19)

$$d f(x;v) = f'(x;v) = f^\circ(x;v) \geqslant \langle\, \zeta, v \rangle \ \ \forall\, v \in \mathbb{R}^n,$$

which implies $\zeta \in \partial_D f(x)$. Hence equality holds throughout in the inclusions.

Conversely, suppose that equality holds in the stated inclusions. Fix $v \in \mathbb{R}^n$. There exists $\zeta \in \partial_C f(x)$ satisfying $f^\circ(x;v) = \langle\, \zeta, v \rangle$. Then $\zeta \in \partial_D f(x)$, whence

$$\langle\, \zeta, v \rangle \leqslant d f(x;v) = \liminf_{t \downarrow 0}\, [\, f(x+tv) - f(x)]/t \leqslant f^\circ(x;v) = \langle\, \zeta, v \rangle.$$

It follows that $f'(x;v)$ exists and agrees with $f^\circ(x;v)$. Thus, f is regular at x. $\square$

Chapter 12

Invariance and monotonicity

A venerable notion from the classical theory of dynamical systems is that of *flow invariance*. When the basic model consists of an autonomous ordinary differential equation $x'(t) = f(x(t))$ and a set S, then flow invariance of the pair (S, f) is the property that for every point $\alpha \in S$, the solution $x(\cdot)$ of the differential equation with initial condition $x(0) = \alpha$ remains in S: $x(t) \in S$ for all $t \geqslant 0$. In this section, we study a highly useful generalization of this concept to situations wherein the differential equation is replaced by an inclusion.

Specifically, let F be a multifunction from $\mathbb{R}^n$ to $\mathbb{R}^n$; that is, $F(x)$ is a subset of $\mathbb{R}^n$ for each x. A **trajectory** x of the multifunction F, on a given interval $[a,b]$, refers to a function $x : [a,b] \to \mathbb{R}^n$ whose n components are absolutely continuous, and which satisfies the differential inclusion

$$x'(t) \in F(x(t)),\ t \in [a,b] \text{ a.e.}$$

When $F(x)$ is a singleton $\{f(x)\}$ for each x, the differential inclusion reduces to an ordinary differential equation. Otherwise, the reader will agree, we would expect there to be *multiple* trajectories from the same initial condition. In such a context, the invariance question bifurcates: do we require *some* of, or *all* of, the trajectories to remain in S? This will be the difference between weak and strong invariance.

Terminology: We extend the notation $\mathrm{AC}[a,b]$, as well as the meaning of the phrase "absolutely continuous" to include vector-valued functions $x : [a,b] \to \mathbb{R}^n$ whose components are absolutely continuous. We shall often refer to such a function as an **arc**, for brevity's sake; sometimes the underlying interval is implicitly defined by the context, as is the dimension n.

F. Clarke, *Functional Analysis, Calculus of Variations and Optimal Control*, Graduate Texts in Mathematics 264, DOI 10.1007/978-1-4471-4820-3_12,

12.1 Weak invariance

Throughout this section, we deal with a multifunction F defined on a given measurable subset Δ of $\mathbb{R}^n$, which may or may not coincide with the set S whose invariance is being studied. The following is always in force:

12.1 Hypothesis. *For every* $x \in \Delta$,

(a) *The set* $F(x)$ *is nonempty, convex, and compact*;

(b) *The graph of* F *is closed at* x:

$$x_i \in \Delta\,,\;\; x_i \to x\,,\;\; v_i \to v\,,\;\; v_i \in F(x_i) \implies v \in F(x)\,;$$

(c) F *is bounded near* x: *there exists* $r = r(x) > 0$ *and* $M = M(x)$ *such that*

$$y \in \Delta \cap B(x,r)\,,\;\; v \in F(y) \implies |v| \leqslant M.$$

We also write the last property in the form $|F(y)| \leqslant M$. We remark that these hypotheses imply that F is a measurable multifunction, by Prop. 6.27.

12.2 Exercise. Prove that (under Hypothesis 12.1, of course) F is bounded on compact sets: for every compact subset C of Δ, there exists $M = M(C)$ such that $|F(x)| \leqslant M \;\; \forall x \in C$. Show, too, that F is *upper semicontinuous* at each $x \in \Delta$:

$$\forall \varepsilon > 0 \; \exists\, r > 0 \text{ such that } y \in B(x,r) \cap \Delta \implies F(y) \subset F(x) + B(0,\varepsilon). \quad \square$$

Another relevant property of F (but imposed only selectively) is that of **linear growth**: this is said to hold on Δ when there exist constants c and d such that

$$x \in \Delta,\; v \in F(x) \implies |v| \leqslant c|x| + d. \tag{1}$$

In light of Exer. 12.2, linear growth holds automatically when Δ is compact.

Existence. The basic result below identifies conditions under which we can assert the existence of a trajectory which remains in a given set S at all times. It turns out, naturally perhaps, that we can achieve this with hypotheses that bear *solely* upon the values that F takes in S itself.

The theorem is framed with the help of the **lower Hamiltonian** h_F associated to F; this is the function defined by

$$h_F(x,\zeta) = \min\big\{\langle \zeta, v\rangle : v \in F(x)\big\}.$$

Note that, in the current context, $h_F : \Delta \times \mathbb{R}^n \to \mathbb{R}$ is finite-valued.

12.3 Theorem. *Let S be closed, and let F satisfy* Hypothesis 12.1 *for $\Delta = S$, together with linear growth. Suppose that*

$$x \in S \implies h_F\big(x, N_S^P(x)\big) \leqslant 0.$$

Then, for any $\alpha \in S$, there is a trajectory x for F defined on $[0,\infty)$ which satisfies

$$x(0) = \alpha, \; x(t) \in S \;\; \forall t \geqslant 0.$$

Remark. The Hamiltonian inequality $h_F(x, N_S^P(x)) \leqslant 0$ in the statement of the theorem means that

$$h_F(x,\zeta) \leqslant 0 \;\; \forall \zeta \in N_S^P(x).$$

Note that this is automatically satisfied at one of the points $\zeta \in N_S^P(x)$, namely the point $\zeta = 0$, since $h_F(x,0) = 0$ by the way h_F is defined. Since $N_S^P(x)$ reduces to $\{0\}$ when $x \in \operatorname{int} S$, the Hamiltonian inequality is automatically satisfied there; thus, only boundary points of S are really involved in considering that inequality.

Concerning the boundary points, the reader will recall that we think of a nonzero proximal normal vector ζ at x as "pointing out" of the set. In this light, the inequality $h_F(x,\zeta) \leqslant 0$ requires that for each such ζ, there be an available velocity direction $v \in F(x)$ such that $\langle \zeta, v \rangle \leqslant 0$. This is very natural, then; in the absence of such v, all trajectories would be forced to leave the set. This is the intuitive interpretation of the Hamiltonian inequality.

Proof.

A. We first prove the theorem under an additional temporary hypothesis [TH] whose removal will be the last step in the proof.

[TH] F is uniformly bounded: $\exists\, M$ such that $F(x) \subset B(0,M) \;\; \forall x \in S$.

We proceed to extend the multifunction F as follows, for any $x \in \mathbb{R}^n$:

$$F(x) = \operatorname{co}\big\{\, v \in F(s) : \; s \in \operatorname{proj}_S(x)\big\}.$$

Note that this agrees with $F(x)$ when $x \in S$. It is easy to verify (with the help of Exer. 12.2) that the set whose convex hull is being taken is compact, and it follows readily that this new F satisfies Hypothesis 12.1 for $\Delta = \mathbb{R}^n$, and, in addition, is globally bounded by M.

Note that any trajectory $x(t)$ for the extended F that satisfies $x(t) \in S \;\; \forall t \geqslant 0$ is a trajectory for the original F. The moral of all this is that, in the presence of [TH], it suffices to prove the theorem when F is globally defined, satisfies Hypothesis 12.1 globally, and is globally bounded by M. We assume this henceforth, until the last step (the removal of [TH]).

B. Let us define a function f as follows: for each x in $\mathbb{R}^n$, choose any $s = s(x)$ in $\operatorname{proj}_S(x)$, and let $v \in F(s)$ be any point minimizing the function $v \mapsto \langle v, x - s\rangle$ over

the set $F(s)$. We set $f(x)=v$. Since $x-s \in N_S^P(s)$ (see Prop. 11.29), this minimum is nonpositive by hypothesis:

$$\langle f(x), x-s(x) \rangle \leqslant 0 \;\; \forall x \in \mathbb{R}^n.$$

By construction, we also have $|f(x)| \leqslant M \;\; \forall x$. We do not claim any other properties for f, such as continuity or even measurability; this will suffice.

Fix any $T>0$, and let

$$\pi_N = \{ t_0, t_1, \ldots, t_{N-1}, t_N \}$$

be a uniform partition of $[0,T]$, where $t_0 = 0$ and $t_N = T$. We proceed by considering, on the interval $[t_0, t_1]$, the differential equation with *constant* right-hand side

$$x'(t) = f(x_0), \quad x(t_0) = x_0 := \alpha.$$

Of course this has a unique affine solution $x(t)$ on $[t_0, t_1]$; we define $x_1 = x(t_1)$. Next, we iterate, by considering on $[t_1, t_2]$ the initial-value problem

$$x'(t) = f(x_1), \quad x(t_1) = x_1.$$

The next so-called *node* of the scheme is $x_2 := x(t_2)$. We proceed in this manner until an arc x_{π_N} (which is in fact piecewise affine) has been defined on all of $[0,T]$; it is usually referred to as the *Euler polygonal arc* corresponding to the partition π_N. Because $|f| \leqslant M$, the polygonal arc x_{π_N} is Lipschitz of rank M, and by Ascoli's theorem, a subsequence of the equicontinuous family x_{π_N} converges uniformly on $[0,T]$ to a Lipschitz arc x.

Let $y_N(t)$ be defined as the node x_i, for the index i in the partition π_N such that $t \in [t_i, t_{i+1})$. Then the y_N are measurable, piecewise constant functions on $[0,T]$ converging uniformly to x. We have

$$x_{\pi_N}{}'(t) \in F_S\big(y_N(t)\big), \; t \in [0,T] \text{ a.e.}$$

Let us prepare the way for an appeal to the weak closure theorem 6.39. Since f is bounded, there is a subsequence of the functions $x_{\pi_N}{}'$ converging in $L^1(0,T)$ to a function w (see Prop. 6.17). Since each x_{π_N} is the indefinite integral of x'_{π_N}, it follows that $w = x'$ a.e. Let us verify hypothesis (b) of the weak closure theorem. We wish to show that for a given $p \in \mathbb{R}^n$, the function $t \mapsto H_{F(u(t))}(p)$ is measurable, for any measurable function $u(\cdot)$. This would follow from the fact that the map $x \mapsto H_{F(x)}(p)$ is Borel measurable. But this is indeed the case, since the hypotheses on F imply that it is upper semicontinuous.

We conclude that in the limit along a subsequence, we obtain $x'(t) \in F(x(t))$ a.e. Thus, x is a trajectory for F. There remains only to prove that $x(t) \in S \;\; \forall t \in [0,T]$, for then the theorem follows (under [TH], still) by iterating the current step on the interval $[T, 2T]$, etc.

C. Let x_{π_N} be one of the sequence of polygonal arcs constructed above converging uniformly to x on $[0,T]$. As before, we denote its node at t_i by x_i $(i = 0,1,\ldots,N)$; thus, $x_0 = x(0) = \alpha$. There exists for each i a point $s_i \in \text{proj}_S(x_i)$ such that $\langle f(x_i), x_i - s_i \rangle \leqslant 0$, as we saw earlier. Note that $s_0 = \alpha$. We calculate

$$\begin{aligned} d_S^2(x_1) &\leqslant |x_1 - s_0|^2 \quad (\text{since } s_0 \in S) \\ &= |x_1 - x_0|^2 + |x_0 - s_0|^2 + 2\langle x_1 - x_0, x_0 - s_0 \rangle \\ &\leqslant M^2(t_1 - t_0)^2 + d_S^2(x_0) + 2\int_{t_0}^{t_1} \langle x'_{\pi_N}(t), x_0 - s_0 \rangle\, dt \\ &= M^2(t_1 - t_0)^2 + d_S^2(x_0) + 2\int_{t_0}^{t_1} \langle f(x_0), x_0 - s_0 \rangle\, dt \\ &\leqslant M^2(t_1 - t_0)^2 + d_S^2(x_0). \end{aligned}$$

The same estimates at any node apply to give

$$d_S^2(x_i) \leqslant d_S^2(x_{i-1}) + M^2(t_i - t_{i-1})^2,$$

whence

$$\begin{aligned} d_S^2(x_i) &\leqslant d_S^2(x_0) + M^2 \textstyle\sum_{\ell=1}^{i} (t_\ell - t_{\ell-1})^2 \\ &\leqslant d_S^2(\alpha) + (TM^2/N) \textstyle\sum_{\ell=1}^{i} (t_\ell - t_{\ell-1}) \leqslant (TM)^2/N. \end{aligned}$$

Now consider the subsequence of x_{π_N} which converges uniformly to x. Since the last estimate holds at every node, we deduce in the limit $d_S(x(t)) \leqslant 0 \;\; \forall\, t \in [0,T]$. Thus the trajectory x remains in S. This completes the proof of the theorem when [TH] holds.

D. Instead of [TH], we assume now only the linear growth condition (1), in which we can suppose that $c > 0$; we set $T_0 = 1/(2c)$. Fix any point α in S, and proceed to choose $R > 0$ so that

$$\frac{R}{cR + c|\alpha| + d} = T_0.$$

(The reader will note that this is possible.) Let us observe that for any point x in $S \cap B(\alpha, R)$, for any $v \in F(x)$, we have

$$|v| \leqslant c|x| + d \leqslant c|x - \alpha| + c|\alpha| + d \leqslant cR + c|\alpha| + d =: M.$$

We redefine F by setting

$$F(x) = \text{co}\,\big\{ F(s) : s \in \text{proj}_{S \cap B(\alpha,R)}(x) \big\}.$$

Note that the new F agrees with the old on $S \cap B(\alpha, R)$. It is easy to see that the redefined F satisfies Hypothesis 12.1 on S, and, in addition, is globally bounded by M (that is, [TH] holds).

Accordingly, we may apply the case of Theorem 12.3 proved above to find a trajectory x on $[0,\infty)$ for the redefined F which lies in S, and which satisfies $x(0)=\alpha$. But for $t\leqslant R/M=T_0$, we necessarily have $x(t)\in B(\alpha,R)$; it follows that on $[0,T_0]$, the arc x is a trajectory for the original F itself.

The theorem statement now follows by iteration: we proceed to find, as above, a trajectory on $[T_0,2T_0]$ starting at $x(T_0)$ which lies in S, and so on. The essential point here is that T_0 does *not* depend on the initial condition. □

The role of linear growth. In the last step of the proof above, the linear growth condition was invoked in order to obtain an *a priori* estimate for the trajectory. It is well known (in some circles, at least), that in the absence of this condition, what is called "finite-time blowup" can occur. This is the phenomenon in which the solution $x(t)$ of a differential equation fails to be defined beyond a certain point at which $|x(t)|$ tends to $+\infty$. A classic example ($n=1$) is the (unique) solution to the Cauchy problem

$$x'(t)=x(t)^2+1,\ \ x(0)=0,$$

which happens to be the function $x(t)=\tan t$. Then x is defined only for $t<\pi/2$; the interval of definition is intrinsically restricted. This phenomenon does *not* happen when the right side of the differential equation has linear growth (as in the case of a linear differential equation), as can be seen from Gronwall's lemma (Theorem 6.41). It is often convenient to postulate linear growth in order to avoid having to mention maximal intervals of definition.

Lurking within Theorem 12.3 is an existence theorem for differential inclusions, one that makes no reference to a set S.

12.4 Corollary. *For a given $\alpha\in\mathbb{R}^n$, let F satisfy* Hypothesis 12.1 *for $\Delta=B(\alpha,r)$, where $r>0$. Then there exists a trajectory x for F on an interval $[0,T]$, $T>0$, satisfying $x(0)=\alpha$. If F satisfies* Hypothesis 12.1 *for $\Delta=\mathbb{R}^n$, as well as (global) linear growth, we can assert that the trajectory is defined on $[0,\infty)$.*

Proof. We seek to apply Theorem 12.3, with $S=B(\alpha,r)$. The trick is to redefine F on the boundary of this ball so that the proximal criterion holds, while retaining Hypothesis 12.1. The proximal criterion is certainly satisfied if the new F contains 0. Accordingly, we redefine $F(x)$ (when $x\in\partial S$) to be the set

$$F(x)=\operatorname{co}\{F(x)\cup\{0\}\}.$$

Then Theorem 12.3 applies, and yields a trajectory for the new F which, until it reaches the boundary of the ball, is also a trajectory for the original F.

When the strengthened global conditions hold, we concatenate trajectory pieces of fixed length to obtain one defined on $[0,\infty)$, as in the proof of Theorem 12.3. □

The property of admitting a trajectory that evolves in the set S turns out to be an important one, meriting a definition.

12.5 Definition. *The system* (S,F) *is called* **weakly invariant** *provided that, for any* $\alpha \in S$*, there exists* $T > 0$ *and a trajectory* x *for* F *on* $[0,T]$ *such that*

$$x(0) = \alpha,\ x(t) \in S\ \forall t \in [0,T].$$

Note that we speak of this property (which is also called *viability*) as being one of the *pair* (S,F), and not just of S. Since F is autonomous (has no dependence on t), the choice of $t = 0$ as the initial time of the trajectories is simply a convenient one that has no intrinsic meaning.

We remark that in contrast to Theorem 12.3, the trajectories here are not necessarily defined on all of $\mathbb{R}_+$. Thus, there is a local nature to the property (the T above depends on α). However, when linear growth holds, it follows as in Cor. 12.4 that weak invariance can be characterized by globally defined trajectories:

12.6 Corollary. *Let* S *be closed, and let* F *satisfy* Hypothesis 12.1 *for* $\Delta = S$*, as well as linear growth on* S*. Then* (S,F) *is weakly invariant if and only if for any* α *in* S*, there exists a trajectory* x *for* F *on* $[0,\infty)$ *such that* $x(0) = \alpha,\ x(t) \in S\ \forall t \geqslant 0$.

In the absence of linear growth, weak invariance remains a purely local property. This is reflected by the local nature of the characterizations given below.

12.7 Theorem. (Weak invariance criteria) *Let* S *be closed, and let* F *satisfy* Hypothesis 12.1 *for* $\Delta = S$*. Then the following are equivalent*:

(a) (S,F) *is weakly invariant.*

(b) $F(x) \cap T_S(x) \neq \emptyset\ \forall x \in S$.

(c) $F(x) \cap \operatorname{co} T_S(x) \neq \emptyset\ \forall x \in S$.

(d) $h_F\big(x, N_S^P(x)\big) \leqslant 0\ \forall x \in S$.

(e) $h_F\big(x, N_S^L(x)\big) \leqslant 0\ \forall x \in S$.

Remark. The list may seem excessively long, but we assure the reader that there is a point to it. To *rule out* weak invariance, for example, it suffices to exhibit a single point $x \in S$ for which $F(x) \cap T_S(x) = \emptyset$; for this purpose, (b) is better than (c). But to *verify* weak invariance, it is easier to check (c) rather than (b). Similar remarks hold concerning the normal cone criteria.

Proof. We have (see Prop. 11.26)

$$N_S^P(x) \subset N_S(x) = T_S(x)^{\triangle} = [\operatorname{co} T_S(x)]^{\triangle}.$$

It follows easily that (b) $\Rightarrow$ (c) $\Rightarrow$ (d). A simple limiting argument shows that (d) and (e) are equivalent (this uses the fact that F is locally bounded). Let us now prove that (a) implies (b).

Accordingly, let us suppose (a), and fix $\alpha \in S$. There is a trajectory x on an interval $[0,T]$ such that $x(0)=\alpha$ and $x(t) \in S\ \ \forall t \geqslant 0$. Since F is bounded near α, x is Lipschitz on some interval $[0,\delta]$, $\delta > 0$. We may therefore choose a sequence t_i decreasing to 0 such that $v := \lim\,(x(t_i)-x(0))/t_i$ exists. Note that $v \in T_S(\alpha)$ by definition.

Now fix $\varepsilon > 0$. By the graph-closedness and local boundedness of F, there exists $\tau > 0$ such that

$$F(x(t)) \subset F(x(0))+B(0,\varepsilon)\ \ \forall t \in [0,\tau]$$

(see Exer. 12.2). Now let us observe that

$$\frac{x(t_i)-x(0)}{t_i} \;=\; \frac{1}{t_i}\int_0^{t_i} x'(s)\,ds.$$

For all i sufficiently large, the integrand above has values almost everywhere in $F(x(s))$, which is contained in $F(x(0))+B(0,\varepsilon)$. Since this last set is convex, we deduce from Exer. 2.44 (for such i):

$$\frac{x(t_i)-x(0)}{t_i} \;\in\; F(x(0))+B(0,\varepsilon).$$

Passing to the limit, we derive $v \in F(x(0))+B(0,\varepsilon)$. Since $\varepsilon > 0$ is arbitrary, we have $v \in F(x(0)) = F(\alpha)$, and (b) follows.

We complete the proof of the theorem by showing that (d) implies (a). Fix any $\alpha \in S$. There exists $r > 0$ and M such that

$$x \in S \cap B(\alpha,r),\ v \in F(x) \implies |v| \leqslant M.$$

We redefine F by setting

$$F(x) \;=\; \operatorname{co}\big\{ F(s) : s \in \operatorname{proj}_{S \cap B(\alpha,r)}(x) \big\}.$$

Note that the new F agrees with the old on $S \cap B(\alpha,R)$. It is easy to see that the redefined F satisfies Hypothesis 12.1 on S, and, in addition, is globally bounded by M (and so, exhibits linear growth). We may therefore apply Theorem 12.3 to find a trajectory x on $[0,\infty)$ for the redefined F which lies in S, and which satisfies $x(0)=\alpha$. But for $t \leqslant r/M =: T$, we necessarily have $x(t) \in B(\alpha,R)$; it follows that, on $[0,T]$, x is a trajectory for the original F; this establishes (a). □

12.8 Corollary. *Let F satisfy* Hypothesis 12.1 *on S, and suppose that at each point $x \in \partial S$, either $F(x) \cap \operatorname{co} T_S(x) \neq \emptyset$, or else $h_F(x,\zeta) \leqslant 0\ \forall \zeta \in N_S^P(x)$. Then (S,F) is weakly invariant.*

Proof. In view of the theorem, it suffices to prove that $\operatorname{co} T_S(x) \subset [N_S^P(x)]^{\triangle}$, for then the proximal criterion holds at every x. The required fact follows from taking polars in $N_S^P(x) \subset N_S(x) = [T_S(x)]^{\triangle} = [\operatorname{co} T_S(x)]^{\triangle}$. □

The point of the corollary is that we can apply different criteria at different points x (tangential or normal, the weaker one in each case), depending upon which one is more convenient at that point. For example, suppose that $N_S^P(x)$ reduces to $\{0\}$ at a certain x. Then there is no need to examine $T_S(x)$, since $h_F(x,0) = 0$; at other points, it may be easier to do so.

Remark. It is clear that in using the tangential or normal criteria of the theorem, only boundary points need be examined, since, when x lies in the interior of S, we have $T_S(x) = \mathbb{R}^n$ and $N_S^P(x) = \{0\}$, in which case all the criteria are trivially satisfied. The tangential criteria only make sense in S, of course, since the tangent cone is undefined outside the set. The proximal criterion, however, can be verified "from the outside," in a sense that we now explain.

Suppose that, for any $y \notin S$ and $x \in \operatorname{proj}_S(y)$, there exists $v \in F(x)$ such that $\langle v, y-x \rangle \leqslant 0$. Since all proximal normals are generated this way (Prop. 11.29), it follows that condition (d) of Theorem 12.7 holds, so that S is weakly invariant. This observation motivates us to define $f(y) = v$, and to conjecture that the differential equation $y' = f(y)$ has solutions that "move towards S."

Making this vague statement precise is beyond the scope of our present discussion, in part because f will not be continuous in general (so a new solution concept for the differential equation needs to be introduced). But this *proximal aiming* technique has been useful in feedback design and stabilization.

12.9 Example. Let $S = \{x \in \mathbb{R}^n : f_i(x) \leqslant 0,\ i = 1, 2, \ldots, k\}$ be a manifold with boundary, as studied in Cor. 10.44. Thus, each f_i is continuously differentiable. We wish to characterize the weak invariance of S with respect to the trajectories of a multifunction F which satisfies Hypothesis 12.1 as well as linear growth on $\mathbb{R}^n$. In doing so, we shall suppose that the functional description of S is nondegenerate, which in this context means that at every boundary point x of S, the active constraints defining S are positively linearly independent.

Then Cor. 10.44 describes the tangent and normal cones to S at x. We deduce, using the tangential criterion, that (S,F) is weakly invariant if and only if, for every x in ∂S, there exists $v \in F(x)$ such that $\langle f_i'(x), v \rangle \leqslant 0\ \ \forall\, i \in I(x)$. The normal criterion, for its part, is easily seen to amount to the inequality

$$\max_{\lambda \in \mathbb{R}^n_+} \ \min_{v \in F(x)} \Big\langle \sum\nolimits_{i \in I(x)} \lambda_i f_i'(x), v \Big\rangle \leqslant 0 .$$

Invoking the minimax theorem 4.36 in order to switch the max and the min, it follows (exercise) that this condition is equivalent to the existence of $v \in F(x)$ such that $\langle f_i'(x), v \rangle \leqslant 0\ \ \forall\, i \in I(x)$. Thus, we obtain the same criterion as the tangential one identified above (as expected). □

12.2 Weakly decreasing systems

We now turn to a *functional* counterpart of weak invariance, one that plays an important role in control theory and differential equations. It turns out to subsume and extend weak invariance, and to be useful in the study of such topics as Lyapunov functions and viscosity solutions.

Let Ω be an open subset of $\mathbb{R}^n$. An arc x such that $x(0) \in \Omega$ is said to be *maximally defined* relative to Ω if either:

- x is defined on $[0,\infty)$ and $x(t) \in \Omega \;\; \forall t \geqslant 0$, or
- x is defined on a finite interval $[0,T]$, and satisfies

$$x(t) \in \Omega \;\; (0 \leqslant t < T), \;\; x(T) \in \partial\Omega.$$

The T of the second case is referred to as the *exit time*, and we denote it by $T(x,\Omega)$. In the first case above, we set $T(x,\Omega) = +\infty$.

We remark that when F satisfies Hypothesis 12.1 on $\Delta = \text{cl}\,\Omega$ as well as linear growth, then, given any $\alpha \in \Omega$, there is a trajectory x for F that is maximally defined relative to Ω and which satisfies $x(0) = \alpha$. This is a consequence of Cor. 12.4. The question we now address is whether we can find such a trajectory that has the additional property of not increasing the value of a given function φ. This odd question turns out to be of real interest in several settings.

Let $\varphi : \mathbb{R}^n \to \mathbb{R}_\infty$ be a given function, and let F be a multifunction defined on an open subset Ω of $\mathbb{R}^n$. We say that (φ, F) is **weakly decreasing** in Ω provided that, for every $\alpha \in \Omega \cap \text{dom}\,\varphi$, there exists a trajectory x for F, maximally defined relative to Ω, such that

$$x(0) = \alpha, \;\; \varphi\big(x(t)\big) \leqslant \varphi(\alpha) \;\; \forall t \in \big[0, T(x,\Omega)\big).$$

When $\Omega = \mathbb{R}^n$, the use of indicator functions and epigraphs reveals the close link between weak decrease and weak invariance.

12.10 Exercise. Let F satisfy Hypothesis 12.1 for $\Delta = \mathbb{R}^n$, as well as linear growth.

(a) Let S be a subset of $\mathbb{R}^n$. Show that (I_S, F) is weakly decreasing in $\mathbb{R}^n$ if and only if (S, F) is weakly invariant.

(b) Let $\varphi : \mathbb{R}^n \to \mathbb{R}_\infty$ be given. Show that (φ, F) is weakly decreasing in $\mathbb{R}^n$ if and only if $(\text{epi}\,\varphi, F \times \{0\})$ is weakly invariant. □

Because weak decrease requires that the trajectory be maximally defined, it is not an entirely local concept, in contrast to weak invariance. For this reason, linear growth is a natural hypothesis to retain in studying weak decrease.

12.11 Theorem. *Let F satisfy* Hypothesis 12.1 *on $\Delta = \operatorname{cl}\Omega$, as well as linear growth. Let $\varphi : \mathbb{R}^n \to \mathbb{R}_\infty$ be lower semicontinuous. Then (φ, F) is weakly decreasing in Ω if and only if*

$$h_F\big(x, \partial_P\,\varphi(x)\big) \leqslant 0 \;\; \forall x \in \Omega\,.$$

Remark. In keeping with that of Theorem 12.3, the notation $h_F(x, \partial_P\,\varphi(x)) \leqslant 0$ is understood to mean $h_F(x, \zeta) \leqslant 0 \;\forall\, \zeta \in \partial_P\,\varphi(x)$, a condition which is automatically satisfied when $\partial_P\,\varphi(x)$ is empty or when $\varphi(x) = +\infty$. As before, the proximal criterion has a natural interpretation. In the fully smooth case, the derivative $d/dt\,\varphi(x(t))$ (which, for decrease, should be nonpositive) equals $\langle \nabla\varphi(x(t)), x'(t)\rangle$; thus, we want $x'(t)$ to be a value $v \in F(x(t))$ for which this inner product is nonpositive, whence the proximal inequality. The inequality, incidentally, is a nonsmooth version of the classical *Hamilton-Jacobi* inequality that arises in many different contexts.

Proof.

A. We redefine F on the boundary of Ω as follows. For $x \in \partial\Omega$, we set

$$F_0(x) = \operatorname{co}\big\{\,\{0\} \cup F(x)\,\big\}.$$

It follows easily that F_0 continues to satisfy Hypothesis 12.1, as well as linear growth, on $\operatorname{cl}\Omega$. We observe that the weak decrease property of (φ, F) is unaffected by this redefinition, so there is no loss of generality in proving the theorem for F_0.

B. Consider now the system (S, F_+), where

$$S = \big(\operatorname{cl}\Omega \times \mathbb{R}\big) \cap \operatorname{epi}\varphi\,,\;\; F_+(x,y) = F_0(x) \times \{0\}\,,\;\; (x,y) \in \mathbb{R}^n \times \mathbb{R}\,.$$

Lemma 1. *The weak decrease of (φ, F) in Ω is equivalent to the weak invariance of the system (S, F_+).*

Proof. Let (φ, F) be weakly decreasing in Ω, and let $(\alpha, r) \in S$. (Thus, $\varphi(\alpha) \leqslant r$.) If $\alpha \in \partial\Omega$, then $F_+(\alpha, r)$ contains $(0,0)$, so that the constant function $(x(t), y(t))$ given by (α, r) is a trajectory for F_+ that remains in S.

Otherwise, if $\alpha \in \Omega$, there is a maximally defined trajectory x for F such that $\varphi(x(t)) \leqslant \varphi(\alpha) \leqslant r$ for $t \in [0, T(x,\Omega))$. The function $(x(t), r)$ is a trajectory for F_+ that lies in S for all $t \in [0, T(x,\Omega))$; we have proved that (S, F_+) is weakly invariant.

Now suppose that (S, F_+) is weakly invariant. Then by Theorem 12.7, we have

$$h_{F_+}(x, y, \zeta, -\lambda) \leqslant 0 \;\; \forall\, (x,y) \in S\,,\;\; (\zeta, -\lambda) \in N_S^P(x,y)\,.$$

Let $\alpha \in \Omega \cap \operatorname{dom}\varphi$ be given; then the point $(\alpha, \varphi(\alpha))$ lies in S. Since F_+ satisfies linear growth, we may deduce from Theorem 12.3 the existence of a trajectory (x, y)

for F_+ on $[0,\infty)$ which satisfies

$$(x(0),y(0)) = (\alpha,\varphi(\alpha)),\ (x(t),y(t)) \in S\ \forall t \geqslant 0.$$

It follows that, for all $t \geqslant 0$, we have $y(t) = \varphi(\alpha)$, $\varphi(x(t)) \leqslant \varphi(\alpha)$. The arc $x(t)$ is a trajectory for F on the interval $[0,T(x,\Omega))$, and as such is maximally defined relative to Ω. It therefore confirms the weak decrease property for (φ,F).

C. We now invoke Theorem 12.7, in the light of Lemma 1: The weak decrease of (φ,F) in Ω is equivalent to the condition

$$h_{F_+}(x,y,\zeta,-\lambda) \leqslant 0\ \forall (x,y) \in S,\ (\zeta,-\lambda) \in N_S^P(x,y).$$

The Hamiltonian inequality is automatically satisfied at points (x,y) for which $x \in \partial\Omega$, since $F_+(x,y)$ contains $(0,0)$ at such points. We may therefore limit the preceding condition to $x \in \Omega$, obtaining one that is fully equivalent:

$$h_F(x,\zeta) \leqslant 0\ \forall (\zeta,-\lambda) \in N^P_{\operatorname{epi}\varphi}(x,r),\ \ \forall (x,r) \in \operatorname{epi}\varphi \cap \big(\Omega \times \mathbb{R}\big). \tag{1}$$

To prove the theorem, it now suffices to establish:

Lemma 2. *The condition* (1) *and the following condition* (2) *are equivalent*:

$$h_F(x,\zeta) \leqslant 0\ \forall \zeta \in \partial_P \varphi(x)\ \ \forall x \in \Omega. \tag{2}$$

Assume first that (1) holds, and let $\zeta \in \partial_P\varphi(x)$, where $x \in \Omega$. Then, by Theorem 11.32, we have $(\zeta,-1) \in N^P_{\operatorname{epi}\varphi}(x,\varphi(x))$. It follows from (1) that $h_F(x,\zeta) \leqslant 0$, which verifies (2).

Now assume that (2) holds, and let $(\zeta,-\lambda) \in N^P_{\operatorname{epi}\varphi}(x,r)$, $x \in \Omega$. We wish to show that $h_F(x,\zeta) \leqslant 0$, so we may assume $\zeta \neq 0$.

If $\lambda > 0$, then $r = \varphi(x)$ necessarily, and we have $\zeta/\lambda \in \partial_P\varphi(x)$ by Theorem 11.32. Then, invoking the hypothesis for $\partial_P\varphi(x)$, we have $h_F(x,\zeta/\lambda) \leqslant 0$, whence $h_F(x,\zeta) \leqslant 0$ by positive homogeneity.

If $\lambda = 0$, then, by Theorem 11.31, there exist sequences x_i and ζ_i, and a corresponding positive sequence λ_i, such that

$$(\zeta_i,-\lambda_i) \to (\zeta,0),\quad (\zeta_i,-\lambda_i) \in N^P_{\operatorname{epi}\varphi}\big(x_i,\varphi(x_i)\big),\quad (x_i,\varphi(x_i)) \to (x,\varphi(x)).$$

By the preceding argument (since $\lambda_i > 0$), we have $h_F(x_i,\zeta_i) \leqslant 0$. Thus there exists $v_i \in F(x_i)$ such that $\langle \zeta_i, v_i \rangle \leqslant 0$. Since F is locally bounded (in view of linear growth), the sequence v_i is bounded, and we may suppose $v_i \to v$; by Hypothesis 12.1, we have $v \in F(x)$. But then $\langle \zeta, v \rangle \leqslant 0$, whence $h_F(x,\zeta) \leqslant 0$. □

12.12 Exercise. Prove that, under the hypotheses of Theorem 12.11, (φ,F) is weakly decreasing in Ω if and only if $h_F(x,\partial_L\varphi(x)) \leqslant 0\ \ \forall x \in \Omega$. (Thus, the subdifferential criterion can be phrased in terms of $\partial_L\varphi$.) □

12.3 Strong invariance

The system (S,F) is said to be **strongly invariant** if it is weakly invariant, and, in addition, *every* trajectory x for F on an interval $[0,T]$ $(T>0)$ which has $x(0)\in S$ satisfies $x(t)\in S\ \forall t\in[0,T]$. It turns out that this property, unlike weak invariance, cannot be assured by purely "internal" hypotheses that hold only within S itself, of the type that are listed in Theorem 12.7, unless the behavior of F is strengthened. The following example illustrates the phenomenon.

12.13 Exercise. Let $f:\mathbb{R}\to\mathbb{R}$ be the function $f(x)=x^{1/3}$.

(a) Show that there are three distinct arcs $x(t)$ on $[0,1]$ of the form ct^p that solve the ordinary differential equation $x'(t)=f(x(t))$. (One of these is evident: $x\equiv 0$.)

(b) Let $S=\{0\}$ and let F be the multifunction $F(x)=\{f(x)\}$. Observe that F satisfies Hypothesis 12.1. Show that the system (S,F) is weakly, but not strongly, invariant.

The point here is that, even though $F(0)\subset T_S(0)$, some trajectories of F nonetheless manage to escape S. □

Note that the (weak) quantifier "there exists" in the definition of weak invariance (Def. 12.5) is replaced in the definition of strong invariance by the (strong) quantifier "every." Correspondingly, in the theorem below, it will be a *strong* Hamilton-Jacobi inequality (a maximum bounded above) that characterizes the property, as opposed to the weak inequality $h_F(x,\zeta)\leqslant 0$ (a minimum bounded above). The inequality in question will be formulated with the *upper Hamiltonian* H_F, the function defined by

$$H_F(x,\zeta)\ =\ \max\big\{\langle\zeta,v\rangle : v\in F(x)\big\}.$$

As explained above, the characterization of strong invariance requires an additional hypothesis on F, a Lipschitz property that will also be important later in a different context.

12.14 Definition. *The multifunction F is said to be Lipschitz on a set C if $F(x)$ is defined and nonempty for every $x\in C$ and, for some constant K, we have*

$$x,y\in C,\ v\in F(x)\implies \exists\, w\in F(y):\ |v-w|\leqslant K|x-y|,$$

or, equivalently, if $x,y\in C\implies F(x)\subset F(y)+K|x-y|B$. A multifunction F is said to be Lipschitz near x if there is a neighborhood V of x such that F is Lipschitz on V.

In characterizing strong invariance, the list of tangential and normal criteria is even longer than that of Theorem 12.7 (for weak invariance), since T_S^C and N_S^C may now join the fray. The reader may as well see the result.

12.15 Theorem. (Strong invariance criteria) *Let S be closed. Let F satisfy* Hypothesis 12.1 *on S, and let F be Lipschitz near every point in ∂S. Then the following are equivalent*:

(a) $H_F\big(x, N_S^C(x)\big) \leqslant 0 \;\; \forall x \in S$

(b) $F(x) \subset T_S^C(x) \;\; \forall x \in S$

(c) $F(x) \subset T_S(x) \;\; \forall x \in S$

(d) $F(x) \subset \operatorname{co} T_S(x) \;\; \forall x \in S$

(e) $H_F\big(x, N_S^P(x)\big) \leqslant 0 \;\; \forall x \in S$

(f) $H_F\big(x, N_S^L(x)\big) \leqslant 0 \;\; \forall x \in S$

(g) (S,F) *is strongly invariant.*

Note: the Lipschitz hypothesis requires that F be defined on a somewhat larger set than S itself.

Proof.

A. Because $N_S^C(x)$ and $T_S^C(x)$ are polars of one another, the equivalence of (a) and (b) is immediate. That (b) $\Longrightarrow$ (c) $\Longrightarrow$ (d) is evident, since $T_S^C(x) \subset T_S(x)$. From

$$\operatorname{co} T_S(x) \subset [T_S(x)]^{\triangle\triangle} = [N_S(x)]^{\triangle} \subset N_S^P(x)^{\triangle},$$

we deduce that (d) $\Longrightarrow$ (e). That (e) implies (f) follows from a simple limiting argument.

B. We now address the implication (f) $\Longrightarrow$ (g). When (f) holds, the system (S,F) is weakly invariant, by Theorem 12.7. Now let x be any trajectory for F on an interval $[0,T]$ such that $x(0) = \alpha \in \partial S$. We undertake to show that, for some $\varepsilon \in (0,T]$, we have $x(t) \in S \;\; \forall t \in [0,\varepsilon]$; this is easily seen to imply strong invariance.

Hypothesis 12.1 implies that, for some $r > 0$ and M, we have $|F(y)| \leqslant M$ for all $y \in B(\alpha,r)$. We may also take r sufficiently small so that F is Lipschitz on the ball $B(\alpha,r)$, with constant K. There exists $\varepsilon \in (0,T]$ such that

$$t \in [0,\varepsilon],\, s \in \operatorname{proj}_S\big(x(t)\big) \implies x(t) \in B(\alpha,r),\, s \in B(\alpha,r).$$

It follows that x is Lipschitz (with constant M) on the interval $[0,\varepsilon]$. We proceed to define $f(t) = d_S(x(t))$; observe that f is Lipschitz on $[0,\varepsilon]$.

Lemma. $f'(t) \leqslant K f(t),\, t \in (0,\varepsilon)$ a.e.

Proof. Let $\tau \in (0,\varepsilon)$ be such that $f'(\tau)$ exists, $x'(\tau)$ exists, and $x'(\tau) \in F(x(\tau))$ (almost all points in $(0,\varepsilon)$ satisfy these conditions). We show that $f'(\tau) \leqslant K f(\tau)$. If $f(\tau) = 0$, then f attains a minimum at τ, whence $f'(\tau) = 0$, and the required inequality holds. We may assume, then, that $f(\tau) > 0$. Let $s \in \operatorname{proj}_S(x(\tau))$. Then, by the closest point characterization of proximal normals (see Prop. 11.29), we have

$$\zeta := (x(\tau) - s)/|x(\tau) - s| \in N_S^P(s).$$

By the Lipschitz condition for F on $B(\alpha,r)$, there exists $v \in F(s)$ such that

$$|v - x'(\tau)| \leqslant K|x(\tau) - s|.$$

Then, using (f), and bearing in mind that ζ is a unit vector, we derive

$$\langle \zeta, x'(\tau) \rangle \leqslant \langle \zeta, v \rangle + \langle \zeta, x'(\tau) - v \rangle \leqslant H_F(s, \zeta) + K|x(\tau) - s| \leqslant K|x(\tau) - s|.$$

This leads to

$$\begin{aligned} f'(\tau) &= \lim_{\delta \to 0} \frac{d_S\big(x(\tau+\delta)\big) - d_S\big(x(\tau)\big)}{\delta} \leqslant \lim_{\delta \to 0} \frac{|x(\tau+\delta) - s| - |x(\tau) - s|}{\delta} \\ &= \langle \zeta, x'(\tau) \rangle \leqslant K|x(\tau) - s| = Kf(\tau), \end{aligned}$$

and the lemma is proved. □

Since f is Lipschitz, nonnegative, and 0 at 0, it follows from the lemma, together with Gronwall's lemma (Theorem 6.41), that f is identically zero on $[0, \varepsilon]$, completing the proof that (f) $\Longrightarrow$ (g).

C. We now show that (g) $\Longrightarrow$ (e), limiting ourselves to points in the boundary of S. Consider any $x \in \partial S$, and fix any v in $F(x)$. There is a ball $B(x, r)$, $r > 0$, upon which F is bounded and Lipschitz.

Let $f(u)$ be the closest point in $F(u)$ to v. It is easy to prove that f is continuous on a ball $B(x, \rho)$, for some $\rho \in (0, r)$. Note that $f(x) = v$. We define a multifunction $\widetilde{F}(u)$ that equals $\{f(u)\}$ for $u \in B(x, \rho)$, and otherwise equals

$$\operatorname{co}\big\{ w : w \in F(u),\ u \in B(x, \rho) \big\}.$$

We also define the closed set $\widetilde{S} = (S \cap B(x, \rho)) \bigcup (\mathbb{R}^n \backslash B^\circ(x, \rho))$.

It is routine to check that $\widetilde{F}$ satisfies Hypothesis 12.1 globally, as well as linear growth. Then trajectories of $\widetilde{F}$, defined on $[0, \infty)$, exist from any initial condition, by Cor. 12.4. The strong invariance of (S, F) implies that the system $(\widetilde{S}, \widetilde{F})$ is strongly, and hence weakly, invariant.

Letting $\tilde{h}$ be the lower Hamiltonian of $\widetilde{F}$, we deduce from Theorem 12.7 that, for any $\zeta \in N_S^P(x)$:

$$\tilde{h}(x, \zeta) = \langle \zeta, f(x) \rangle = \langle \zeta, v \rangle \leqslant 0.$$

Since v is arbitrary in $F(x)$, (e) follows.

D. To complete the proof of the theorem, it suffices to prove that (e) $\Longrightarrow$ (b). Let x lie in ∂S, and let $v \in F(x)$; we need to show that v belongs to $N_S^L(x)^\triangle = T_S^C(x)$.

Now any element ζ of $N_S^L(x)$ is of the form $\zeta = \lim_i \zeta_i$, where $\zeta_i \in N_S^P(x_i)$ and $x_i \to x$. Letting K be a Lipschitz constant for F in a neighborhood of x, there is, for each i sufficiently large, a point $v_i \in F(x_i)$ such that $|v_i - v| \leqslant K|x_i - x|$, and (by (e)) we have $\langle \zeta_i, v_i \rangle \leqslant 0$. We obtain in the limit $\langle \zeta, v \rangle \leqslant 0$ as required. □

In similar manner to Cor. 12.8, we have:

12.16 Corollary. *In the context of the theorem, suppose that at each $x \in \partial S$, we have either $F(x) \subset \operatorname{co} T_S(x)$, or else $H_F(x, N_S^P(x)) \leqslant 0$. Then the system (S,F) is strongly invariant.*

Strongly decreasing systems. Let $\varphi : \mathbb{R}^n \to \mathbb{R}_\infty$ be a given function. We say that the system (φ, F) is *strongly decreasing* in an open set Ω provided that, for every α in $\Omega \cap \operatorname{dom}\varphi$, there exists a trajectory x for F, maximally defined relative to Ω, such that $x(0) = \alpha$ and

$$s,t \in [0, T(x,\Omega)),\ s \leqslant t \implies \varphi(x(t)) \leqslant \varphi(x(s)),$$

and provided this monotonicity holds for *every* maximally defined trajectory.

12.17 Theorem. *Let F satisfy* Hypothesis 12.1 *on $\Delta = \operatorname{cl}\Omega$, as well as linear growth. Let F be locally Lipschitz in Ω, and let $\varphi : \mathbb{R}^n \to \mathbb{R}_\infty$ be lower semicontinuous. Then (φ, F) is strongly decreasing in Ω if and only if*

$$H_F(x,\zeta) \leqslant 0 \ \ \forall\, \zeta \in \partial_P \varphi(x) \ \ \forall\, x \in \Omega\,.$$

Proof. As in the proof of Theorem 12.11, and for the same auxiliary system (S, F_+) as defined there, it follows that

Lemma 1. *The strong decrease of (φ, F) in Ω is equivalent to the strong invariance of the system (S, F_+).*

The corresponding Hamiltonian equivalence is established in somewhat the same fashion as before, but the Lipschitz condition now plays a role.

Lemma 2. *The following are equivalent*:

$$H_F(x,\zeta) \leqslant 0 \ \ \forall\, \zeta \in \partial_P \varphi(x),\ \ \forall\, x \in \Omega \tag{1}$$

$$H_F(x,\zeta) \leqslant 0 \ \ \forall\, (\zeta,-\lambda) \in N^P_{\operatorname{epi}\varphi}(x,r),\ \ \forall\, (x,r) \in \operatorname{epi}\varphi \cap \big(\Omega \times \mathbb{R}\big) \tag{2}$$

To prove this, assume first that (2) holds, and let $\zeta \in \partial_P \varphi(x)$, where $x \in \Omega$. Then, by Theorem 11.32, we have $(\zeta,-1) \in N^P_{\operatorname{epi}\varphi}(x,\varphi(x))$. It follows from (2) that $H_F(x,\zeta) \leqslant 0$, which verifies (1).

Now assume that (1) holds, and let $(\zeta,-\lambda) \in N^P_{\operatorname{epi}\varphi}(x,r)$, where $x \in \Omega$. We wish to show that $H_F(x,\zeta) \leqslant 0$. Clearly we may suppose $\zeta \neq 0$. Fix any $v \in F(x)$; we wish to prove $\langle \zeta, v \rangle \leqslant 0$.

If $\lambda > 0$, then $r = \varphi(x)$ necessarily (see Exer. 11.30), and we have $\zeta/\lambda \in \partial_P \varphi(x)$ by Theorem 11.32. Then, by the hypothesis for $\partial_P \varphi(x)$, we have $H_F(x, \zeta/\lambda) \leqslant 0$, whence $\langle \zeta, v \rangle \leqslant 0$ by positive homogeneity.

If $\lambda = 0$, then, by Theorem 11.31, there exist sequences x_i and ζ_i, and a corresponding positive sequence λ_i, such that

$$(\zeta_i, -\lambda_i) \to (\zeta, 0), \quad (\zeta_i, -\lambda_i) \in N^P_{\operatorname{epi}\varphi}\big(x_i, \varphi(x_i)\big), \quad (x_i, \varphi(x_i)) \to (x, \varphi(x)).$$

By the preceding argument (since $\lambda_i > 0$), we have $H_F(x_i, \zeta_i) \leqslant 0$. There exists $v_i \in F(x_i)$ such that $|v_i - v| \leqslant K|x_i - x|$, by the Lipschitz condition for F. But then $\langle \zeta_i, v_i \rangle \leqslant 0$, whence $\langle \zeta, v \rangle \leqslant 0$, as required. □

Lyapunov functions. The reader will encounter applications of system monotonicity later, in connection with generalized solutions of the Hamilton-Jacobi equation, and sufficient conditions in optimal control. We pay a brief visit here to another of its applications, the theory of Lyapunov functions.

For purposes of illustration, let the multifunction F satisfy Hypothesis 12.1 on $\mathbb{R}^n$, as well as linear growth. Suppose that $\varphi : \mathbb{R}^n \to \mathbb{R}_+$ is continuous and satisfies

$$\varphi(x) > 0 \implies h_F\big(x, \partial_P \varphi(x)\big) \leqslant -\omega,$$

where ω is a positive constant. Then the very existence of such a function φ says something about the *controllability* of the system described by F: from any initial condition, there is a trajectory that is driven to the set $\{\varphi = 0\}$ (which turns out to be nonempty necessarily).

12.18 Proposition. *For any $\alpha \in \mathbb{R}^n$, there is a trajectory x for F with $x(0) = \alpha$ such that, for some $T \leqslant \varphi(\alpha)/\omega$, we have $x(T) \in \varphi^{-1}(0)$.*

Proof. We can assume $\varphi(\alpha) > 0$. For an augmented state $(x,y) \in \mathbb{R}^n \times \mathbb{R}$, consider the system data

$$\Omega = \big(\mathbb{R}^n \backslash \varphi^{-1}(0)\big) \times \mathbb{R}, \;\; F_+(x,y) = F(x) \times \{1\}, \;\; \varphi_+(x,y) = \varphi(x) + \omega y.$$

Then F_+ satisfies Hypothesis 12.1 as well as linear growth. Note that any element (ζ, θ) of $\partial_P \varphi_+(x,y)$ is of the form (ζ, ω), where $\zeta \in \partial_P \varphi(x)$. It follows that

$$h_{F_+}(x,y,\zeta,\theta) \leqslant 0 \;\; \forall (x,y) \in \Omega, \;\; \forall (\zeta,\theta) \in \partial_P \varphi_+(x,y).$$

We may therefore apply Theorem 12.11 to these data, with the choice of initial condition $\alpha_+ := (\alpha, 0) \in \Omega$. We obtain a trajectory x for F with $x(0) = \alpha$, maximally defined with respect to $\mathbb{R}^n \backslash \varphi^{-1}(0)$, such that

$$\varphi(x(t)) + \omega t \leqslant \varphi(\alpha) \;\; \forall t \in [0, T),$$

where T is the exit time from $\mathbb{R}^n \backslash \varphi^{-1}(0)$. Since φ is nonnegative, it follows that $T \leqslant \varphi(\alpha)/\omega$, for otherwise we obtain $\varphi(x(t)) < 0$, a contradiction. □

In the best known case of the above, the founding one, φ is taken to be smooth (on $\mathbb{R}^n \backslash \{0\}$) and *positive definite* (that is, $\varphi(0) = 0$ and $\varphi(x) > 0 \;\forall x \neq 0$), and $F(x)$ is of the form $\{f(x)\}$, where f is a continuous function. Then Prop. 12.18 asserts that, when such a φ exists, the differential equation $x' = f(x)$ is *globally asymptotically stable*. This means that globally defined solutions exist from any initial condition,

and all of them converge to the zero set of φ, namely $\{0\}$. The function φ is referred to as a *Lyapunov function*.

This celebrated technique to verify stability was introduced in 1900 by Lyapunov. Note that in this setting, H_F and h_F coincide, and strong and weak coalesce: stability equals controllability.[1]

In the general case, when F is not a singleton, controllability is a different issue, and induces a number of interesting questions of its own. One of these is the following: given a Lyapunov function φ, taken to be smooth for simplicity, how can we use it to design a *feedback* mechanism that has the effect of steering x to the origin? This refers to a selection f of F with the property that the solutions of the differential equation $x' = f(x)$ approach the set $\{\varphi = 0\}$ in a suitable sense.

The obvious attempt is to choose $f(x) \in F(x)$ to be a "steepest descent" direction, in order to obtain the decrease rate $\langle \nabla\varphi(x), f(x)\rangle \leqslant -\omega$. However, this leads to *discontinuous* functions f in general, which complicates matters in an intriguing fashion; but this is another story...

[1] It is of interest to know under what conditions the existence of a Lyapunov function φ is necessary, as well as sufficient, for system stability; we shall not pursue the issue of such *converse Lyapunov theorems*, however.

Chapter 13

Additional exercises for Part II

13.1 Exercise. Let M be an $n \times n$ symmetric matrix. Consider the constrained optimization problem

$$\text{Minimize } \langle x, Mx \rangle \text{ subject to } h(x) := 1 - |x|^2 = 0, \; x \in \mathbb{R}^n.$$

(a) Observe that a solution x_* exists, and write the conclusions of the multiplier rule (Theorem 9.1) for this problem. Show that they cannot hold abnormally. It follows that the multiplier in this case is of the form $(1, \lambda)$.

(b) Deduce that λ is an eigenvalue of M, and that $\lambda = \langle x_*, Mx_* \rangle$.

(c) Prove that λ coincides with the first, or least eigenvalue λ_1 of M (this statement makes sense because the eigenvalues are real). Deduce the *Rayleigh formula*, which asserts that λ_1 is given by $\min \{ \langle x, Mx \rangle : |x| = 1 \}$. □

13.2 Exercise. The *linear-quadratic* optimization problem is that of minimizing

$$f(x) = \langle x, Qx \rangle + \langle b, x \rangle$$

over $x \in \mathbb{R}^n$, and under the constraint $Ax = c$. Here, the given data consist of points $b \in \mathbb{R}^n$, $c \in \mathbb{R}^k$, a symmetric $n \times n$ matrix Q, and a $k \times n$ matrix A (with $k \leqslant n$). A is assumed to have maximal rank.

(a) Write the necessary conditions of Theorem 9.4 for this problem, and verify that they can hold only in normal form.

We suppose henceforth that Q is positive definite.

(b) Prove the existence of a solution x_* to the problem.

(c) Show that the $k \times k$ matrix $AQ^{-1}A^T$ is invertible and positive definite.

(d) Show that the unique solution x_* is given by

F. Clarke, *Functional Analysis, Calculus of Variations and Optimal Control*,
Graduate Texts in Mathematics 264, DOI 10.1007/978-1-4471-4820-3_13,

$$x_* = \frac{1}{2}\,Q^{-1}\Big\{A^T\big(AQ^{-1}A^T\big)^{-1}\big(2c+AQ^{-1}b\big)-b\Big\}.$$

(e) Find the unique multiplier λ. What information does it contain? □

13.3 Exercise. Show that $(1,1)$ is the projection of $(2,2)$ onto the convex set

$$\big\{(x,y) : x^2+2y^2 \leqslant 3,\; x^2-y \leqslant 0\big\}.$$ □

13.4 Exercise. We consider the problem

$$\min \int_0^1 \Lambda(t,v(t))\,dt \;:\; v \in L^\infty(0,1),\; v(t) \geqslant 0 \text{ a.e.},\; \int_0^1 v(t)\,dt \leqslant q.$$

We suppose that the function $(t,v) \mapsto \Lambda(t,v) \in \mathbb{R}$ is continuous in t and convex in v, and that $q > 0$. Prove that the admissible function v_* is a minimizer for the problem if and only if there exists $\gamma \in R_+$ such that

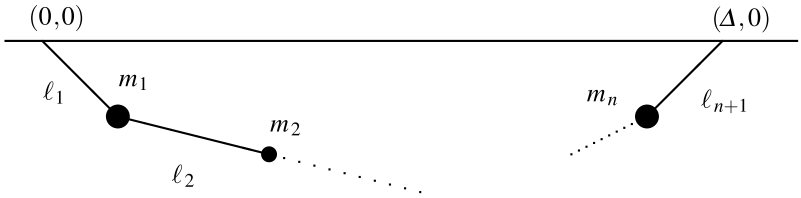
$$-\gamma \in \partial_v \Lambda(t,v_*(t)) + N_{[0,\infty)}(v_*(t)), \quad t \in [0,1] \text{ a.e.}$$ □

13.5 Exercise. A system consists of n masses $m_1, m_2, \ldots, m_n$ joined by cords of lengths $\ell_1, \ell_2, \ldots, \ell_{n+1}$, and hanging from two fixed points $(0,0)$ and $(\Delta,0)$, where $\Delta > 0$; see the figure below. It is assumed that the mass of the cord is negligible.

(a) Let x_1 be the (signed) vertical difference (or change in height) between $(0,0)$ and the first mass m_1, x_2 between m_1 and m_2, etc., up to x_n. Let x_{n+1} be the vertical difference between the last mass m_n and the point $(\Delta,0)$. Thus, $x_1 < 0$ and $x_{n+1} > 0$.

Now define $w_i = \sum_i^n m_j$ (for $1 \leqslant i \leqslant n$) and $w_{n+1} = 0$. Show that the equilibrium position (the one that minimizes potential energy) corresponds to finding the $x \in \mathbb{R}^{n+1}$ minimizing the function $\sum_1^{n+1} w_i x_i$ subject to the constraints

$$-\ell_i \leqslant x_i \leqslant \ell_i, \quad \sum\nolimits_1^{n+1} x_i = 0, \quad \sum\nolimits_1^{n+1} \sqrt{\ell_i^2 - x_i^2} = \Delta\,.$$



We define

$$f(x) = \sum_1^{n+1} w_i x_i\,,\ h(x) = \sum_1^{n+1} x_i\,,\ g(x) = \Delta - \sum_1^{n+1} \sqrt{\ell_i^2 - x_i^2}$$
$$S = \{x \in \mathbb{R}^{n+1} : -\ell_i \leqslant x_i \leqslant \ell_i\,,\ i = 1, 2, \ldots, n+1\}$$

(b) With these data, show that the problem of finding the equilibrium has the form of the problem (P) considered in §9.2. Observe that the constraint $g(x) = 0$ has been replaced by an inequality; what justifies this change? Note that with this modification, the problem lies within the convex case.

(c) Prove that the problem admits a solution $\bar{x}$, provided that $\sum_i \ell_i \geqslant \Delta$.

(d) Write the necessary conditions of Theorem 9.4, and show that the Slater condition is satisfied if and only if $\sum_i \ell_i > \Delta$ (which we assume henceforth).

(e) If it so happens that for some $i \in \{1, 2, \ldots, n\}$, we have

$$\Big|\sum_{j=1}^{i-1} \ell_j - \sum_{j=i+1}^{n+1} \ell_j\Big| \leqslant \ell_i\,,$$

then show that the solution is given by

$$\bar{x}_k = \begin{cases} -\ell_k & \text{if } 1 \leqslant k \leqslant i-1 \\ \sum_{j=1}^{i-1} \ell_j - \sum_{j=i+1}^{n+1} \ell_j & \text{if } k = i \\ \ell_k & \text{if } i+1 \leqslant k \leqslant n+1 \end{cases}$$

This is the case in which the system hangs with some slack in the cord that joins m_{i-1} and m_i, each mass being as low as it could possibly be. We rule it out by supposing henceforth that $|\bar{x}_i| < \ell_i \ (i = 1, 2, \ldots, n+1)$.

(f) Show that the multiplier γ corresponding to g is nonzero.

(g) Prove the existence of $\gamma > 0$ and λ such that

$$\bar{x}_i = \frac{-(w_i + \lambda)\,\ell_i}{\sqrt{(w_i + \lambda)^2 + \gamma^2}}\,,\quad i = 1, 2, \ldots, n+1.$$

Thus, the solution $\bar{x}$ is completely determined by the two scalar multipliers γ and λ. We now turn to duality to suggest how these multipliers might be found.

(h) Show that the dual problem consists of maximizing on $\mathbb{R}^2$ the function

$$(\gamma, \lambda) \mapsto \Delta\gamma - \sum_i \ell_i \sqrt{(w_i + \lambda)^2 + \gamma^2}$$

subject to $\gamma \geqslant 0$. (Note that this problem has two unknowns, regardless of n.) It is clear that the resulting multiplier λ is negative; why is this consistent with its sensitivity interpretation? □

13.6 Exercise. Solve the problem described in Exer. 6.12. □

13.7 Exercise. (Fenchel duality) Let $f : X \to \mathbb{R}_\infty$ and $g : X \to \mathbb{R}_\infty$ be convex and bounded below, where X is a normed space. We posit the existence of a point x_0 in $\operatorname{dom} f \cap \operatorname{dom} g$ at which g is continuous. The goal is to prove

$$\inf_{x \in X} f(x) + g(x) = \max_{\zeta \in X^*} -f^*(\zeta) - g^*(-\zeta).$$

(Note the use of "max" on the right side.) Without loss of generality, we may take $x_0 = 0$. Define $V : X \to \mathbb{R}_\infty$ by

$$V(\alpha) = \inf_{x \in X} f(x) + g(x - \alpha).$$

(a) Show that

$$V(0) = \inf_{x \in X} f(x) + g(x) \geqslant \sup_{\zeta \in X^*} -f^*(\zeta) - g^*(-\zeta).$$

(b) Prove that V is convex, and bounded above in a neighborhood of 0.

(c) Prove the existence of $\zeta \in \partial V(0)$.

(d) Show that $-f^*(\zeta) - g^*(-\zeta) \geqslant V(0)$, and conclude. □

13.8 Exercise. Let S be a nonempty, closed, and convex subset of a normed space X. Show that for any $y \in X$ we have

$$d_S(y) = \max\left\{ \langle \zeta, y \rangle - H_S(\zeta) : \|\zeta\|_* \leqslant 1 \right\},$$

where H_S is the support function of S. □

13.9 Exercise. The following problem arises in signal processing. It consists of finding $x \in X = L^2(-\pi, \pi)$ which minimizes

$$\int_{-\pi}^{\pi} |x(t)|\, dt$$

under the constraints

$$E(x) \leqslant E_0,\ F_k(x) = c_k\ (k = 1, 2, \dots, n),$$

where

$$E(x) = \int_{-\pi}^{\pi} |x(t)|^2\, dt,\quad F_k(x) = \frac{1}{\sqrt{\pi}} \int_{-\pi}^{\pi} (\cos kt)\, x(t)\, dt.$$

(We seek the best estimate of the bounded energy signal displaying certain known Fourier coefficients.)

(a) Prove the existence of a solution x_*.

(b) Write the necessary conditions of Theorem 9.4, justifying their use.

(c) We assume henceforth $\sum_1^n c_k^2 < E_0$; prove then that the Slater condition holds.

(d) We assume henceforth $\sum_1^n c_k^2 > 0$; prove then that $E(x_*) = E_0$. (Hint: show that the multiplier associated with the constraint is nonzero.)

(e) Prove the existence of $\gamma_* > 0$ and $\lambda_* \in \mathbb{R}^n$ such that the solution x_* is given, for almost every t, by

$$x_*(t) = \frac{-1}{2\sqrt{\pi}\,\gamma_*}\Big[\,|p(t)| - \sqrt{\pi}\,\Big]_+ \frac{p(t)}{|p(t)|},$$

where

$$p(t) := \sum_{k=1}^n \lambda_{*_k} \cos(kt),$$

and where $[q]_+$ means $\max[q,0]$. (Note that the formula for $x_*(t)$ is naturally interpreted as yielding 0 when $p(t) = 0$.) [Hint: Theorem 6.31 is useful here.]

(f) Formulate explicitly the dual problem whose solution is (γ_*, λ_*). □

13.10 Exercise. Let $f : \mathbb{R} \to \mathbb{R}$ be given by

$$f(x) = |x|^{3/2} \sin\left(\frac{1}{\sqrt{|x|}}\right),$$

where we set $f(0) = 0$. Show that f is locally Lipschitz and differentiable, with $f'(0) = 0$. Prove that $\partial_P f(0) = \emptyset$ and $\partial_C f(0) = [-1/2, 1/2]$. (The question of which of the three sets $\partial_P f(0)$, $\{f'(0)\}$, or $\partial_C f(0)$ best reflects the nature of f near 0 is open to debate.) □

13.11 Exercise. Find the tangent cones T and T^C, as well as the normal cones N, N^L and N^C, for each of the following subsets of $\mathbb{R}^2$, at the origin:

$$S_1 = \{(x,y) : xy = 0\} \qquad S_2 = \{(x,y) : y \geqslant 2|x|\}$$

$$S_3 = \text{cl}\{\mathbb{R}^2 \backslash S_2\} \qquad S_4 = \{(x,y) : y \leqslant \sqrt{|x|}\}. \qquad □$$

13.12 Exercise. Let f be the function of Example 10.28.

(a) Calculate $\partial_C f_1(0)$ and $\partial_C f_2(0)$, where $f_1(x) = f(x,0)$ and $f_2(y) = f(0,y)$.

(b) Show that $\partial_C f_1(0) \times \partial_C f_2(0) \not\subset \partial_C f(0,0) \not\subset \partial_C f_1(0) \times \partial_C f_2(0)$. □

13.13 Exercise. Let $f : X \to \mathbb{R}$ and $g : X \to \mathbb{R}$ be regular at x, where X is a Banach space. Prove that $\max(f,g)$ is regular at x. □

13.14 Exercise. (Eigenvalue design) Certain optimal design problems involving discrete structures or electrical networks are phrased in terms of the eigenvalues of an $n \times n$ symmetric matrix $M = M(y)$ depending on a design parameter $y \in \mathbb{R}^m$. To be specific, suppose that each entry m_{ij} of M is a smooth function of y (and that

$m_{ij}(y) = m_{ji}(y)\ \forall i,j$). Set $\Lambda(y)$ equal to the maximal eigenvalue of $M(y)$. Then the object might be to minimize Λ, subject perhaps to certain constraints on y. Prove that the function $y \mapsto \Lambda(y)$ is locally Lipschitz and regular.

In general, we expect the function Λ to be nonsmooth and nonconvex. Verify that such is the case in the following example ($n = m = 2$):

$$M(y_1, y_2) = \begin{bmatrix} y_1^2 & y_2^2 - 1 \\ y_2^2 - 1 & y_1^2 \end{bmatrix}.$$ □

13.15 Exercise. In the context of the multiplier rule of Theorem 10.47, suppose in addition that the functions f, g, and h are continuously differentiable. Prove that the stationarity condition may be expressed as follows:

$$\big\{\eta f + \langle \gamma, g\rangle + \langle \lambda, h\rangle\big\}'(x_*; v) \geqslant 0 \ \forall v \in T_S^C(x_*).$$ □

13.16 Exercise. Let $f : \mathbb{R}^m \times \mathbb{R}^n \to \mathbb{R}$ be a continuous function such that, for each $u \in \mathbb{R}^m$, the function $g_u : \mathbb{R}^n \to \mathbb{R}$ defined by $g_u(v) = f(u,v)$ is convex. Prove that

$$(\theta, \zeta) \in \partial_L f(u,v) \implies \zeta \in \partial g_u(v).$$ □

13.17 Exercise. We establish the following refinement of Cor. 11.7.

Proposition. *Suppose that $f : \mathbb{R}^n \to \mathbb{R}_\infty$ is lsc, and that $x \in \operatorname{dom} f$. Let there be positive constants ε, K such that*

$$u \in B(x, \varepsilon),\ f(u) < f(x) + \varepsilon,\ \zeta \in \partial_P f(u) \implies |\zeta| \leqslant K.$$

Then f is Lipschitz of rank K near x.

(a) Let $g(u) := \min\big[f(u), f(x) + \varepsilon/2\big]$. Show that if ζ belongs to $\partial_P g(u)$, then either $\zeta \in \partial_P f(u)$ with $f(u) < f(x) + \varepsilon$, or else $\zeta = 0$.

(b) Use this fact to prove that g is Lipschitz near x.

(c) Deduce that f Lipschitz near x. □

13.18 Exercise. Let $f : U \to \mathbb{R}$ be continuous, where $U \subset \mathbb{R}^n$ is open and convex. Let S_i ($i = 1, 2, \dots, k$) be subsets of U satisfying

$$\operatorname{cl} U = \operatorname{cl}\big\{\textstyle\bigcup_{i=1}^k S_i\big\},$$

and suppose that, for some constant K, the function f restricted to S_i is Lipschitz of rank K for each i. Prove that f is Lipschitz of rank K. □

13.19 Exercise. Let S be a nonempty closed subset of $\mathbb{R}^n$, and let $x \notin S$. Prove that

$$\partial_L d_S(x) = \big\{(x - s)/|x - s| : s \in \operatorname{proj}_S(x)\big\}.$$ □

13.20 Exercise. Prove the following theorem, a proximal version of the decrease principle (Theorem 5.22):

Theorem. *Let $f : \mathbb{R}^n \to \mathbb{R}_\infty$ be lower semicontinuous. Suppose that for some x in* $\operatorname{dom} f$ *and positive numbers δ and r we have*

$$u \in B^\circ(x,r) \cap \operatorname{dom} f,\ \ \zeta \in \partial_P f(u) \implies |\zeta| \geqslant \delta.$$

Then

$$\inf \big\{ f(u) : u \in B^\circ(x,r) \big\} \leqslant f(x) - \delta r.$$

□

13.21 Exercise. With the help of Exer. 13.20, and taking $X = \mathbb{R}^n$, prove that Theorem 5.31 continues to hold if Hypothesis 5.30 is replaced by the following one, which does not require the existence of directional derivatives: There is an open neighborhood V of $(\bar{x}, \bar{y})$ and $\delta > 0$ with the following property:

$$(x,y) \in V,\ \varphi(x,y) > 0,\ \zeta \in \partial_P \varphi(x,y) \implies |\zeta| \geqslant \delta.$$

Here, the proximal subdifferential is taken with respect to the x variable. (We continue to assume that φ is continuous.) □

13.22 Exercise.

(a) Find the tangent and normal cones at $(0,0,0)$ to each of the following sets:

$$S_1 = \big\{ (x,y,z) \in \mathbb{R}^3 : x^3 - y + z^2 = 0,\ x = y \big\},$$
$$S_2 = \big\{ (x,y,z) \in \mathbb{R}^3_+ : x - y + z \leqslant x^2 \big\}.$$

(Why is there no ambiguity about which types of tangents or normals are involved?)

(b) Show that if $(a,b,c,d) \in N^L_E(0,0,0,0)$, where E is the set

$$E = \big\{ (x,y,z,u) \in \mathbb{R}^4 : x + y + |u| \geqslant 0,\ x = z \big\},$$

then $a = b - c$, $b \leqslant 0$, and $d = \pm b$. □

13.23 Exercise. Let $\Lambda : \mathbb{R}^n \to \mathbb{R}$ be convex. Fix $v_* \in \mathbb{R}^n$ and define

$$f(\alpha) = \Lambda(v_*/\alpha)\,\alpha,\ \ \alpha > 0.$$

Prove that f is convex on $(0, +\infty)$, and that if $\gamma \in \partial f(1)$, then there exists an element ζ in $\partial \Lambda(v_*)$ such that $\gamma = \Lambda(v_*) - v_* \bullet \zeta$. □

13.24 Exercise. The exercise establishes a measurability property of the multifunctions $\partial_L f$ and $\partial_C f$. We consider a function $f : \mathbb{R} \times \mathbb{R}^n \to \mathbb{R}$ such that

- $t \mapsto f(t,x)$ is measurable for each $x \in \mathbb{R}^n$.
- $x \mapsto f(t,x)$ is locally Lipschitz for each $t \in \mathbb{R}$.

The notation $\partial_L f(t,x)$ below (for example) refers to the L-subdifferential of the function $f(t,\cdot)$ evaluated at x, for t fixed. The goal is to prove

Proposition. *Let $u : \mathbb{R}^n \to \mathbb{R}$ be measurable. Then the multifunctions*

$$t \mapsto \partial_L f\big(t,u(t)\big) \quad \textit{and} \quad t \mapsto \partial_C f\big(t,u(t)\big)$$

are measurable.

Define, for any positive integer i, the function

$$\theta_i(t,x,\zeta) = \min\big\{ f(t,x+y) - f(t,x) - \langle \zeta, y\rangle + i|y|^2 : |y| \leqslant 1/i \big\}.$$

(a) Show that $\theta_i : \mathbb{R}\times\mathbb{R}^n\times\mathbb{R}^n \to \mathbb{R}$ is measurable with respect to t and locally Lipschitz with respect to x and ζ, and prove that

$$\zeta \in \partial_P f(t,x) \iff \exists\, i \geqslant 1 \text{ such that } \theta_i\big(t,x,\zeta\big) = 0.$$

(b) Prove that for any compact subset V of $\mathbb{R}^n$, the following set is measurable:

$$\big\{ t : \partial_L f\big(t,u(t)\big) \cap V \neq \emptyset \big\},$$

and deduce the proposition. □

13.25 Exercise. The Dini derivate and subdifferential may be used to define tangent and normal cones. Given a closed subset S of $\mathbb{R}^n$, we could, for example, use its indicator function I_S in order to define

$$T_S^D(x) = \big\{ v \in \mathbb{R}^n : dI_S(x;v) \leqslant 0 \big\},\ N_S^D(x) = \partial_D I_S(x).$$

(The use of the distance function d_S instead of the indicator function would yield the same constructs.) Show that we obtain in this way exactly the classical tangent and normal cones $T_S(x)$ and $N_S(x)$ of § 1.4 (p. 20). □

13.26 Exercise. Let $f : \mathbb{R}^2 \to \mathbb{R}$ be defined by

$$f(x,y) = \max\big\{ \min(x,y), -[\max(0,-x)]^{3/2} \big\}.$$

Prove that at $(0,0)$, the sets $\partial_P f$, $\partial_D f$, $\partial_L f$, and $\partial_C f$ are all different. Building upon this, find a closed subset E of $\mathbb{R}^3$ containing $(0,0,0)$ such that, at the origin, the cones N_E^P, N_E^D $(= N_E$, see Exer. 13.25$)$, N_E^L, and N_E^C are all different. □

13.27 Exercise. The sense in which an arc x solves the differential equation

$$x'(t) = g(x(t)), \quad t \in [a,b] \text{ a.e.}$$

is open to question when the underlying function g fails to be continuous. Requiring pointwise equality almost everywhere, for example, is unsatisfactory: solutions in this sense tend not to exist.

Let $g : \mathbb{R}^n \to \mathbb{R}^n$ be locally bounded, and let us construct a multifunction as follows:

$$G(x) = \bigcap_{\delta>0} \overline{\text{co}}\, g(B(x,\delta)).$$

We may then define a solution of $x'(t) = g(x(t))$ to mean an arc x satisfying the differential inclusion $x'(t) \in G(x(t))$ a.e.[1] Using this solution concept, prove a local existence theorem for the associated Cauchy problem (that is, when $x(a)$ is prescribed). Show that if g happens to be continuous, then a solution in this new extended sense is in fact a solution in the usual sense. □

13.28 Exercise. Invariance can be studied in a non autonomous setting, for example when the set S depends on t, as we illustrate now. Accordingly, let S be a multifunction from $\mathbb{R}_+$ to the subsets of $\mathbb{R}^n$.

We now say that the system (S,F) is *weakly invariant* provided that, for any $\tau \in \mathbb{R}_+$, for any $\alpha \in S(\tau)$, there exists $T > \tau$ and a trajectory x for F on $[\tau,T]$ such that

$$x(\tau) = \alpha,\ x(t) \in S(t) \quad \forall t \in [\tau,T].$$

It is clear that this definition coincides with the earlier one (see Def. 12.5) if S happens to not depend on t. We assume that the graph of S, the set

$$G = \big\{(t,x) \in \mathbb{R}_+ \times \mathbb{R}^n : x \in S(t)\big\},$$

is closed. Our goal is to prove

Theorem. *The system (S,F) is weakly invariant if and only if*

$$(t,x) \in G,\ (\theta,\zeta) \in N_G^P(t,x) \implies \theta + h_F(x,\zeta) \leqslant 0.$$

The proof is based on a device known as "absorbing t into the dynamics." We relabel t as a new initial coordinate x^0 of the state x, and we define $F_+(x^0,x) = \{1\} \times F(x)$.

(a) Prove that (S,F) is weakly invariant if and only if the (autonomous!) system (G,F_+) is weakly invariant.

(b) Apply Theorem 12.7 to conclude the proof, and show that the new theorem reduces to the former one if S does not depend on t.

(c) Extend the other criteria in Theorem 12.7 to the non autonomous setting. □

13.29 Exercise. In a similar vein to the preceding exercise, we study the weak decrease of a system (φ,F) in an open subset Ω of $\mathbb{R}^n$, but in a non autonomous setting. We let F satisfy the same hypotheses as in Theorem 12.11; however, we now let $\varphi : \mathbb{R} \times \mathbb{R}^n \to \mathbb{R}_\infty$ be a lower semicontinuous function that depends on (t,x).

[1] Filippov and Krasovskii have pioneered this type of approach.

We say that (φ, F) is weakly decreasing in Ω provided that, for every (τ, α) in $(\mathbb{R}\times\Omega)\cap \operatorname{dom}\varphi$, there exists a trajectory x for F defined on $[\tau, T(x,\Omega))$ such that

$$x(\tau) = \alpha, \;\; \varphi(t, x(t)) \leqslant \varphi(\tau, \alpha) \;\; \forall t \in [\tau, T(x,\Omega)).$$

Here, $T(x,\Omega)$ is, as before, the exit time from Ω (but now, relative to τ):

$$T(x,\Omega) = \inf\left\{T > \tau : x(T) \in \partial\Omega\right\}.$$

Prove the following:

Theorem. *(φ, F) is weakly decreasing in Ω if and only if*

$$(t,x) \in \operatorname{dom}\varphi, \; (\theta,\zeta) \in \partial_P\varphi(t,x) \implies \theta + h_F(x,\zeta) \leqslant 0.$$

Observe that this reduces to Theorem 12.11 if φ does not depend on t. □

13.30 Exercise. We define the property that (φ, F) is *strongly increasing* by simply reversing the inequality involving φ in the definition of strongly decreasing (p. 270). Under the hypotheses of Theorem 12.17, prove that strong increase is characterized by the following proximal Hamilton-Jacobi inequality:

$$h_F(x,\zeta) \geqslant 0 \;\; \forall \zeta \in \partial_P\varphi(x) \;\; \forall x \in \Omega. \qquad \square$$

13.31 Exercise. The goal is to prove the following result on the existence of zeros.

Theorem. *Let K be a nonempty compact convex subset of $\mathbb{R}^n$, and let $h : K \to \mathbb{R}^n$ be a Lipschitz function satisfying $h(x) \in T_K(x) \; \forall x \in K$. Then there exists $z \in K$ such that $h(z) = 0$.*

(a) Why can we assume that h is globally Lipschitz and bounded on $\mathbb{R}^n$?

(b) Fix $\delta > 0$. By classical results, for any $\alpha \in \mathbb{R}^n$, there is a unique solution $x_\alpha(\cdot)$ defined on $[0, \delta]$ of the Cauchy problem

$$x_\alpha'(t) = h(x_\alpha(t)), \;\; 0 < t < \delta, \; x_\alpha(0) = \alpha.$$

Prove that $\alpha \in K \implies x_\alpha(\delta) \in K$.

(c) Classical theory implies that the function $f_\delta : K \to \mathbb{R}^n$ which maps α to $x_\alpha(\delta)$ is continuous. Thus, in light of the preceding conclusion, we may invoke Brouwer's fixed point theorem to deduce the existence of a fixed point $z_\delta \in K$ of the function f_δ. Complete the proof be letting $\delta \downarrow 0$. □

Generalized Jacobians. Let $F : \mathbb{R}^m \to \mathbb{R}^n$ be locally Lipschitz. The *generalized Jacobian* $\partial F(x)$ is defined as follows:

$$\partial F(x) = \operatorname{co}\left\{\lim_{i\to\infty} DF(x_i) : x_i \to x, \; x_i \notin E \cup E_F\right\},$$

where $E \subset \mathbb{R}^m$ is any set of measure zero, E_F is the set of points at which F fails to be differentiable, and DF is the Jacobian matrix. Then $\partial F(x)$ is a nonempty convex set[2] of $n \times m$ matrices, and is compact when viewed as a subset of $\mathbb{R}^{nm}$.

13.32 Exercise. Prove the two following properties of the generalized Jacobian:

(a) For any $\zeta \in \mathbb{R}^n$, we have $\partial_C \langle \zeta, F \rangle(x) = \zeta^* \partial F(x)$ ($*$ refers here to transpose).

(b) ∂F is graph-closed: $x_i \to x,\ M_i \in \partial F(x_i),\ M_i \to M \implies M \in \partial F(x)$. □

13.33 Exercise. When $n = m$, we say that the generalized Jacobian $\partial F(x)$ defined above is *nonsingular* if each element of $\partial F(x)$ is nonsingular (invertible). The goal is to prove the following (nonsmooth) version of Theorem 5.32:

Theorem. *Let $F : \mathbb{R}^n \to \mathbb{R}^n$ be a mapping which is Lipschitz in a neighborhood of a point $\bar{x}$, and such that $\partial F(\bar{x})$ is nonsingular. Set $\bar{y} = F(\bar{x})$. Then there is a neighborhood U of $(\bar{x}, \bar{y})$ and $K > 0$ such that*

$$d\big(x, F^{-1}(y)\big) \leqslant K|F(x) - y| \quad \forall (x,y) \in U.$$

Proof. Let $\varphi(x,y) = |F(x) - y|$. With the help of the chain rule and Exer. 13.32, prove the existence of a neighborhood V of $(\bar{x}, \bar{y})$ and $\delta > 0$ such that

$$(x,y) \in V,\ \ \varphi(x,y) > 0,\ \zeta \in \partial_P \varphi(x,y) \implies |\zeta| \geqslant \delta,$$

where ∂_P is taken with respect to x. Now use Exer. 13.21 to conclude. □

13.34 Exercise. The goal is to build upon Exer. 13.33 in order to obtain the *Lipschitz inverse function theorem*:

Theorem. (Clarke 1973) *If $F : \mathbb{R}^n \to \mathbb{R}^n$ is Lipschitz near $\bar{x}$ and $\partial F(\bar{x})$ is nonsingular, then there exist a neighborhood W of $F(\bar{x})$, a neighborhood A of $\bar{x}$, and a Lipschitz function $\hat{x} : W \to \mathbb{R}^n$ such that $\hat{x}(F(\bar{x})) = \bar{x}$ and*

$$F(\hat{x}(y)) = y \ \forall y \in W, \quad \hat{x}(F(u)) = u \ \forall u \in A.$$

In order to prove the theorem, establish first that F is locally injective; more specifically, that there exist a neighborhood Ω of x and $\eta > 0$ such that

$$|F(u) - F(x)| \geqslant \eta |u - x| \ \forall u, x \in \Omega.$$

Combine this with Exer. 13.33 in order to complete the proof. □

13.35 Exercise. Show that the theorem of Exer. 13.34 applies to the case

$$n = 2,\ F(x,y) = \big[\,|x| + y,\, 2x + |y|\,\big] \ \text{at } (0,0).$$ □

[2] As in the gradient formula (Theorem 10.27), it can be shown that the definition is independent of the choice of the null set E.

Part III
Calculus of Variations

Chapter 14

The classical theory

We will now discuss in a little more detail the Struggle for Existence.
Charles Darwin (The Origin of Species)

Life is grey, but the golden tree of theory is always green.
Goethe (Journey by Moonlight)

It's the question *that drives us, Neo.*
Morpheus (The Matrix)

The basic problem in the subject that is referred to as the *calculus of variations* consists in minimizing an integral functional of the type

$$J(x) = \int_a^b \Lambda\big(t, x(t), x'(t)\big)\, dt$$

over a class of functions x defined on the interval $[a,b]$, and which take prescribed values at a and b.

The study of this problem (and its numerous variants) is over three centuries old, yet its interest has not waned. Its applications are numerous in geometry and differential equations, in mechanics and physics, and in areas as diverse as engineering, medicine, economics, and renewable resources. It is not surprising, then, that modeling and numerical analysis play a large role in the subject today. In the following chapters, however, we present a course in the calculus of variations which focuses on the core mathematical issues: necessary conditions, sufficient conditions, existence theory, regularity of solutions.

For those like the reader who have a sensitive mathematical conscience, the statement of the basic problem, as rendered above, may well create an uneasiness, a craving for precision. What, exactly, is the class of functions in which x lies? What hypotheses are imposed on the function Λ? Is the integral well defined? Does a solution exist?

In the early days of the subject, these questions went unaddressed, at least explicitly. (Implicitly: everything was taking place in a very smooth universe in which problems evidently had solutions.) Our era, more attuned to the limits of smoothness, requires a more deliberate approach, and a well-defined setting. And this is just as well, for, as the reader will come to understand, the history, the development, and the most useful insights into the subject are inextricably wrapped up with the very questions just posed.

F. Clarke, *Functional Analysis, Calculus of Variations and Optimal Control*, Graduate Texts in Mathematics 264, DOI 10.1007/978-1-4471-4820-3_14,

Hypotheses. The focus of our attention is the integral functional $J(x)$ defined above, where Λ is a function of three variables, and where $[a,b]$ is a given interval in $\mathbb{R}$. $\Lambda(t,x,v)$ is referred to as the **Lagrangian**, and the generic notation for its three variables is (t,x,v): time, state, velocity.

This chapter deals with the case in which these variables are one-dimensional and all the functions involved are smooth. We take $\Lambda : \mathbb{R}^3 \to \mathbb{R}$ to be a twice continuously differentiable function, and we limit attention to functions $x : [a,b] \to \mathbb{R}$ that belong to $C^2[a,b]$. This means that x lies in $C[a,b]$, the derivatives x' and x'' exist and are continuous in (a,b), and both x' and x'' admit continuous extensions to $[a,b]$. It is clear that this is more than adequate to guarantee that the integral defining $J(x)$ is well defined for each competing x.

Given in addition two points A and B in $\mathbb{R}$, we now consider the **basic problem** in the calculus of variations:

$$\text{minimize } J(x) : x \in C^2[a,b],\ x(a) = A,\ x(b) = B. \tag{\textbf{P}}$$

$J(x)$ is referred to as the **cost** corresponding to x. A function $x : [a,b] \to \mathbb{R}$ is termed **admissible** if it satisfies the boundary constraints and lies in the appropriate class, in this case $C^2[a,b]$. A **solution** x_* of (P) refers to an admissible function x_* such that $J(x_*) \leqslant J(x)$ for all other admissible functions x. We also refer to x_* as a **minimizer** for the problem.

14.1 Example. (A minimal surface problem) The well-known problems to which the calculus of variations was first applied arise in geometry and mechanics. A famous example of the problem (P) that goes back to Euler's seminal monograph of 1744 is to find the shape of the curve $x(t)$ joining (a,A) to (b,B) whose associated surface of rotation (about the t-axis) has minimal area.

This can be given a physical interpretation: when a soap surface is spanned by two concentric rings of radius A and B, the resulting surface will be a surface of rotation of a curve $x(t)$, and we expect the area of the surface to be a minimum. This expectation (confirmed by experiment) is based upon *d'Alembert's principle*, which affirms that in static equilibrium, the observed configuration minimizes total potential energy (which, for a soapy membrane desperately seeking to contract, is proportional to its area).

In concrete terms, the soap bubble problem consists of minimizing

$$\int_a^b x(t)\sqrt{1+x'(t)^2}\,dt \quad \text{subject to } x(a) = A,\ x(b) = B.$$

This is the case of the basic problem (P) in which $\Lambda(t,x,v) = x\sqrt{1+v^2}$. (The surface area is in fact given by 2π times the integral, but we can omit this multiplicative factor, which does not effect the minimization.) We shall be seeing this problem again later. □

14.1 Necessary conditions

The following result identifies the first *necessary condition* that a minimizing x must satisfy; it is in effect an analogue of Fermat's rule that $f'(x) = 0$ at a minimum.

Notation: The partial derivatives of the function $\Lambda(t,x,v)$ with respect to x and to v are denoted by Λ_x and Λ_v.

14.2 Theorem. (Euler 1744) *If x_* is a solution of* (P), *then x_* satisfies the* **Euler equation:**

$$\frac{d}{dt}\left\{\Lambda_v\big(t,x_*(t),x_*'(t)\big)\right\} = \Lambda_x\big(t,x_*(t),x_*'(t)\big) \;\; \forall t \in [a,b]. \tag{1}$$

Proof. Euler's proof of this result used discretization, but the now standard proof given here uses Lagrange's idea: a *variation*, from which the subject derives its name. In the present context, a variation means a function $y \in C^2[a,b]$ such that $y(a) = y(b) = 0$. We fix such a y, and proceed to consider the following function g of a single variable:

$$g(\lambda) = J(x_* + \lambda y) = \int_a^b \Lambda\big(t, x_* + \lambda y, x_*' + \lambda y'\big)\,dt. \tag{2}$$

(The reader will notice that we have yielded to the irresistible temptation to leave out certain arguments in the expression for the integral; thus x_* should be $x_*(t)$ and y is really $y(t)$, and so on. Having already succumbed the first time, we shall do so routinely hereafter.) It follows from standard results in calculus that g is differentiable, and that we can "differentiate through the integral" to obtain

$$g'(\lambda) = \int_a^b \left[\Lambda_x\big(t,x_*+\lambda y, x_*'+\lambda y'\big)\,y + \Lambda_v\big(t,x_*+\lambda y, x_*'+\lambda y'\big)\,y'\right]dt. \tag{3}$$

Observe now that for each λ, the function $x_* + \lambda y$ is admissible for (P), whence

$$g(\lambda) = J(x_* + \lambda y) \geqslant J(x_*) = g(0).$$

It follows that g attains a minimum at $\lambda = 0$, and hence that $g'(0) = 0$; thus:

$$\int_a^b \left[\alpha(t)\,y(t) + \beta(t)\,y'(t)\right]dt = 0,$$

where we have set

$$\alpha(t) = \Lambda_x\big(t,x_*(t),x_*'(t)\big), \;\; \beta(t) = \Lambda_v\big(t,x_*(t),x_*'(t)\big).$$

Using integration by parts, we deduce

$$\int_a^b \left[\alpha(t) - \beta'(t)\right] y(t)\, dt \;=\; 0.$$

Since this is true for any variation y, it follows that the continuous function which is the coefficient of y under the integral sign must vanish identically on $[a,b]$ (left as an exercise). But this conclusion is precisely Euler's equation. □

A function $x \in C^2[a,b]$ satisfying Euler's equation is referred to as an **extremal**. The Euler equation (1) is (implicitly) a differential equation of order two for x_*, and one may expect that, in principle, the two boundary conditions will single out a unique extremal. We shall see, however, that it's more complicated than that.

14.3 Exercise.

(a) Show that all extremals for the Lagrangian $\Lambda(t,x,v) = \sqrt{1+v^2}$ are affine. Why is this to be expected?

(b) Show that the Euler equation for the Lagrangian $\Lambda(t,x,v) = x^2 + v^2$ is given by $x'' - x = 0$.

(c) Find the unique admissible extremal for the problem

$$\min \quad \int_0^1 \left(x'(t)^2 + x(t)^2\right) dt \;\;:\;\; x \in C^2[0,1],\; x(0) = 0,\; x(1) = 1. \qquad \square$$

Local minima are extremals. The Euler equation is the first-order necessary condition for the calculus of variations problem (P), and we would expect it to hold for merely *local* minima (suitably defined). We develop this thought now.

A function x_* admissible for (P) is said to provide a **weak local minimum** if, for some $\varepsilon > 0$, for all admissible x satisfying $\|x - x_*\| \leqslant \varepsilon$ and $\|x' - x_*'\| \leqslant \varepsilon$, we have $J(x) \geqslant J(x_*)$. The anonymous norm referred to in a context such as this one will always be that of $L^\infty[a,b]$ (or $C[a,b]$); thus, for example, the notation above refers to

$$\|x - x_*\| \;=\; \max\left\{\, |x(t) - x_*(t)| : t \in [a,b]\,\right\}.$$

The proof of the necessity of Euler's equation goes through for a local minimizer just as it did for a global one: the function g defined in (2) attains a local minimum at 0 rather than a global one; but we still have $g'(0) = 0$, which is what leads to the Euler equation. Thus any weak local minimizer for (P) must be an extremal.

The Erdmann condition. The Lagrangian Λ is said to be **autonomous** if it has no explicit dependence on the t variable. The following consequence of the Euler equation can be a useful starting point in the identification of extremals.

14.4 Proposition. *Let x_* be a weak local minimizer for* (P), *where Λ is autonomous. Then x_* satisfies the* **Erdmann condition:** *for some constant h, we have*

$$x_*'(t)\,\Lambda_v\big(x_*(t),x_*'(t)\big)-\Lambda\big(x_*(t),x_*'(t)\big) = h \;\;\forall\, t\in[a,b].$$

Proof. It suffices to show that the derivative of the function on the left side is zero, which follows from the Euler equation: we leave this as an exercise. □

14.5 Example. (continued) We return to the soap bubble problem (Example 14.1), armed now with some theory. Suppose that x_* is a weak local minimizer for the problem, with $x_*(t) > 0 \;\forall\, t$. The reader may verify that the Euler equation is given by

$$x''(t) = (1+x'(t)^2)/x(t).$$

We deduce from this that x_*' is strictly increasing (thus x_* is strictly convex). Since Λ is autonomous, we may invoke the Erdmann condition (Prop. 14.4). This yields the existence of a positive constant k such that

$$\big(x_*(t)'\big)^2 = x_*(t)^2/k^2 - 1,\;\; t\in[a,b].$$

If x_*' is positive throughout $[a,b]$, we may solve this to reveal the separated differential equation

$$\frac{k\,dx}{\sqrt{x^2-k^2}} = dt\,.$$

Mathematics students used to know by heart a primitive for the left side: the function $k\cosh^{-1}(x/k)$. It follows that x_* is of the form

$$x_*(t) = k\cosh\left(\frac{t+c}{k}\right).$$

A curve of this type is called a *catenary*.[1]

If, instead of being positive, x_*' is negative, a similar analysis shows that, once again, x_* is a catenary:

$$x_*(t) = \kappa\cosh\left(\frac{t+\sigma}{\kappa}\right),$$

for certain constants σ,κ (different from c and k, on the face of it). Since x_*' is strictly increasing, the general case will have x_*' negative, up to a point τ (say), and then positive thereafter. Thus, x_* is a catenary (with constants σ,κ) followed by another catenary (with constants c,k). The smoothness of x_*, it can be shown, forces the constants to coincide, so we can simply assert that x_* is a catenary.

Notice that the detailed analysis of this problem is not a trivial matter. Furthermore, the only rigorous conclusion reached at the moment is the following: if there exists a weak local minimizer $x_*\in C^2[a,b]$ which is positive on $[a,b]$, then x_* is a catenary.

[1] Not every choice of interval $[a,b]$ and prescribed endpoints A,B will define a catenary; we do not pursue this issue, which is carefully analyzed in Bliss [5].

Our impulse may be to accept that conclusion, especially since soap bubbles do demonstrably exist, and have the good grace to cooperate (frequently) by being catenaries. But the example illustrates a general fact in optimization: "real" problems are rarely so impressed by the theory that they immediately reveal their secrets.

An *ad hoc* analysis, sometimes difficult, is often required. In the soap bubble case, for example, note the use of the (presupposed) regularity of x_*, which allowed us to match up the catenaries. (We return to this regularity issue later.) In this text, we stress the theory, rather than the details of particular problems. But we prefer to have warned the reader that no amount of general theory reduces a difficult problem to a simple exercise. □

Our next example involves an important topic in classical mechanics.

14.6 Example. (Least action principle) In 1744, Euler (in that same monograph we mentioned before) extended d'Alembert's principle to mechanical systems which are in motion, rather than in static equilibrium. His celebrated Principle of Least Action[2] postulates that the movement between two time instants t_1 and t_2 minimizes the *action*

$$\int_{t_1}^{t_2} \big(K - V\big)\, dt\,,$$

where K refers to kinetic energy and V to potential energy.

We proceed to illustrate the principle of least action in a simple case: the (unforced) oscillation in the plane of a pendulum of length ℓ whose mass m is entirely in the bob. The angle θ (see Fig. 14.1) is a convenient choice of *generalized coordinate*; the motion of the pendulum is described by the corresponding function $\theta(t)$. In terms of θ, the kinetic energy $K = mv^2/2$ is given by $m\big(\ell\,\theta'\big)^2/2$.

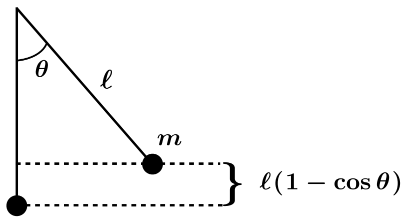


Fig. 14.1
The pendulum

If one uses $\theta = 0$ as the reference level for calculating potential energy mgh, then it is given in terms of θ by $mg\ell(1 - \cos\theta)$, as a little trigonometry shows. Thus the action between two instants t_1 and t_2 is given by

$$\int_{t_1}^{t_2} \Big\{ \tfrac{1}{2}\, m\big(\ell\,\theta'(t)\big)^2 - mg\ell\big(1 - \cos\theta(t)\big)\Big\}\, dt\,.$$

[2] Sometimes mistakenly attributed to Maupertuis; see the discussion in [27].

We apply the least action principle: it follows that the resulting motion $\theta(t)$ satisfies Euler's equation for the action functional. The reader may check that this yields the following differential equation governing the pendulum's movement:

$$\theta''(t) + (g/\ell)\sin\theta(t) = 0.$$

This equation, which can also be deduced from Newton's law, is in fact the one which describes the movement of the pendulum. But does it really minimize the action? Perhaps in a local sense?

Consider the analogy with the minimization of a function $f(x)$ (on $\mathbb{R}$, say). The Euler equation corresponds to the necessary condition $f'(x_*) = 0$, the stationarity of f at a given point x_*. Further evidence of a local minimum would be the second-order condition $f''(x_*) \geqslant 0$. And if we knew in addition that $f''(x_*) > 0$, then we could say with certainty that x_* provides at least a local minimum. In this light, it seems reasonable to pursue second-order conditions in the calculus of variations. The honor of first having done so belongs to Legendre, although, to some extent, he was scorned for his efforts, for reasons that we shall see. □

In studying second-order conditions, and for the rest of this chapter, we strengthen the regularity hypothesis on the Lagrangian by assuming that Λ is C^3.

14.7 Theorem. (Legendre's necessary condition, 1786) *Let x_* be a weak local minimizer for* (P). *Then we have*

$$\Lambda_{vv}\big(t, x_*(t), x_*'(t)\big) \geqslant 0 \;\; \forall\, t \in [a,b].$$

Proof. We consider again the function g defined by (2). We observe that the formula (3) for $g'(\lambda)$ implies that g' is itself differentiable. We proceed to develop an expression for $g''(0)$. Differentiating under the integral in $g'(\lambda)$, and then setting $\lambda = 0$, we obtain

$$g''(0) = \int_a^b \Big[\Lambda_{xx}(t)\,y^2 + 2\Lambda_{xv}(t)\,y\,y' + \Lambda_{vv}(t)\,y'^2\Big]\,dt,$$

where $\Lambda_{xx}(t)$ (for example) is an abbreviation for $\Lambda_{xx}(t, x_*(t), x_*'(t))$, and where we have invoked the fact that Λ_{xv} and Λ_{vx} coincide. We proceed to define

$$P(t) = \Lambda_{vv}\big(t,\, x_*(t),\, x_*'(t)\big) \tag{4}$$

$$Q(t) = \Lambda_{xx}\big(t,\, x_*(t),\, x_*'(t)\big) - \frac{d}{dt}\Lambda_{xv}\big(t,\, x_*(t),\, x_*'(t)\big). \tag{5}$$

(Note that Q is well defined, in part because Λ is C^3.) Using this notation, integration by parts shows that the last expression for $g''(0)$ may be written

$$g''(0) = \int_a^b \Big[P(t)\,y'^2(t) + Q(t)\,y^2(t)\Big]\,dt. \tag{6}$$

Since g attains a local minimum at 0, we have $g''(0) \geqslant 0$. We now seek to exploit the fact that this holds for every variation y. To begin with, a routine approximation argument (see Ex. 21.13) shows that for any Lipschitz (rather than C^2) variation y in $\mathrm{Lip}_0[a,b]$ (the class of Lipschitz functions on $[a,b]$ that vanish at a and b), we still have the inequality (6).

Now let $[c,d]$ be any subinterval of $[a,b]$, and let ε be any positive number. We define a function $\varphi \in \mathrm{Lip}_0[a,b]$ as follows: φ vanishes on $[a,c]$ and $[d,b]$, and, in between (that is, on $[c,d]$), φ is a sawtooth function whose derivative is alternately $+1$ and -1, with the effect that for all $t \in [c,d]$ we have $|\varphi(t)| < \varepsilon$. Then, by taking $y = \varphi$ in (6), we deduce

$$\int_c^d \big[P(t) + |Q(t)|\varepsilon^2\big]\,dt \;\geqslant\; 0.$$

Since $\varepsilon > 0$ is arbitrary, we conclude that the integral of the continuous function $P(t)$ over $[c,d]$ is nonnegative. Since in turn the subinterval $[c,d]$ is arbitrary, we have proved, as required, that P is nonnegative on $[a,b]$. □

14.2 Conjugate points

In contrast to the Euler equation, the Legendre necessary condition has the potential to distinguish between a maximum and a minimum: at a local maximizer x_* of J (that is, a local minimizer of $-J$), we have $\Lambda_{vv}(t, x_*(t), x_*'(t)) \leqslant 0$.

To illustrate the distinction, consider the functional

$$J(x) \,=\, \int_a^b \sqrt{1 + x'(t)^2}\; dt\,.$$

Legendre's condition tells us that it is useless to seek local *maxima* of J, since here we have $\Lambda_{vv} = \{1 + v^2\}^{-3/2} > 0$; only local (or global) minima may exist.

Legendre proceeded to prove (quite erroneously) that his necessary condition, when strengthened to strict inequality, was also a *sufficient* condition for a given extremal to provide a weak local minimum. He was scathingly criticized by Lagrange for his sins.[3]

Legendre's proof went as follows. Using the same function g as above, together with the (Lagrange!) expansion

$$g(1) - g(0) \,=\, g'(0) + (1/2)g''(\lambda) \,=\, (1/2)g''(\lambda)$$

[3] The reader will be relieved to know that Legendre's reputation recovered; his name is one of 72 inscribed on the Eiffel tower...

(the Euler equation corresponds to $g'(0) = 0$), routine calculations (given in the proof of Theorem 14.8 below) lead to the inequality

$$J(x_* + y) - J(x_*) \geqslant \frac{1}{2} \int_a^b \left[(P - \delta) y'^2 + Q y^2 \right] dt \tag{1}$$

for all variations y in a suitable weak neighborhood of 0 (that is, such that $\| y \|$ and $\| y' \|$ are sufficiently small). Here, P and Q have the same meaning as before, and $\delta > 0$ is chosen so that $P(t) - \delta > 0$ on $[a,b]$ (this is where the supposed strict positivity of P is used). There remains only to show, therefore, that the integral term in (1), which we label I, is nonnegative.

To this end, let w be any continuously differentiable function, and note that

$$\begin{aligned} I = \int_a^b \left[(P-\delta) y'^2 + Q y^2 \right] dt &= \int_a^b \left[[(P-\delta) y'^2 + Q y^2 + \big(w y^2\big)' \right] dt \\ = \int_a^b (P-\delta) \Big[y'^2 + \frac{Q + w'}{P - \delta}\, y^2 + 2 \frac{w}{P-\delta}\, y y' \Big] dt \\ = \int_a^b (P-\delta) \Big[y' + \frac{w}{P-\delta}\, y \Big]^2 dt \geqslant 0, \end{aligned}$$

where the factorization in the last integral expression depends upon having chosen the function w (heretofore arbitrary) to satisfy

$$\frac{Q + w'}{P - \delta} = \left(\frac{w}{P - \delta} \right)^2 \iff w' = \frac{w^2}{P - \delta} - Q.$$

The proof appears to be complete. It has a serious defect, however.

Even in the present sophisticated era, students are sometimes surprised that such an innocent-looking differential equation as the one above can *fail* to have a solution w defined on the entire interval $[a,b]$. In fact, the equation is nonlinear, and we know only that a solution exists if b is taken sufficiently close to a, from well-known local existence theorems. In light of this, perhaps we can forgive Legendre his error (taking for granted the existence of w), especially since his approach, suitably adapted, did in fact turn out to be highly fruitful.

We summarize what can be asserted on the basis of the discussion so far.

14.8 Theorem. *Let $x_* \in C^2[a,b]$ be an admissible extremal satisfying the strengthened Legendre condition $\Lambda_{vv}(t, x_*(t), x_*'(t)) > 0 \ \forall t \in [a,b]$. Suppose there exists a function $w \in C^1[a,b]$ satisfying the differential equation*

$$w'(t) = \frac{w(t)^2}{P(t)} - Q(t), \ \ t \in [a,b].$$

Then x_* is a weak local minimizer for (P).

Proof. We begin with the estimate mentioned above.

Lemma. *There is a constant M such that the following inequality holds for every variation y having $\|y\|+\|y'\| \leqslant 1$:*

$$J(x_*+y)-J(x_*) \geqslant \frac{1}{2}\int_a^b \Big[Py'^2+Qy^2\Big]\,dt - M\big\{\|y\|+\|y'\|\big\}\int_a^b y'^2\,dt.$$

To prove this, we first observe that, for some $\lambda \in (0,1)$,

$$J(x_*+y)-J(x_*) = g(1)-g(0) = \frac{1}{2}\,g''(\lambda),$$

by the second-order mean value theorem of Lagrange (also known as the Taylor expansion), since $g'(0) = 0$ (x_* being an extremal). Calculating as we did in the proof of Theorem 14.7, we find

$$g''(\lambda) = \int_a^b \Big[\Lambda^{\lambda}_{vv}(t)\,y'^2 + 2\Lambda^{\lambda}_{xv}(t)\,y\,y' + \Lambda^{\lambda}_{xx}(t)\,y^2\Big]\,dt\,, \tag{2}$$

where, for example, $\Lambda^{\lambda}_{vv}(t)$ is shorthand for

$$\Lambda_{vv}(t, x_*+\lambda y, x_*'+\lambda y').$$

The partial derivatives Λ_{vv}, Λ_{xv} and Λ_{xx}, being continuously differentiable, admit a common Lipschitz constant K on the ball around $(0,0,0)$ of radius

$$|a|+|b|+\|x_*\|+\|x_*'\|+1.$$

This allows us to write, for any variation y as described in the statement of the lemma,

$$\big|\Lambda^{\lambda}_{vv}(t)-\Lambda^{0}_{vv}(t)\big| \leqslant K|\lambda|\,|(y(t),y'(t))| \leqslant K|(y(t),y'(t))|,$$

and similarly for the other two terms in (2). This leads to

$$J(x_*+y)-J(x_*) \geqslant \tfrac{1}{2}\int_a^b \big[\Lambda^{0}_{vv}(t)\,y'^2+2\Lambda^{0}_{xv}(t)\,y\,y'+\Lambda^{0}_{xx}(t)\,y^2\big]\,dt$$
$$-2\big[\|y\|+\|y'\|\big]\int_a^b \big[y'^2+2|y\,y'|+y^2\big]\,dt.$$

The proof of Theorem 14.7 showed that the first term on the right is precisely

$$\frac{1}{2}\int_a^b \big[Py'^2+Qy^2\big]\,dt.$$

To complete the proof of the lemma, therefore, it now suffices to know that for some constant c, we have

$$\int_a^b \left[2|yy'|+y^2\right] dt \leqslant c\int_a^b y'^2\,dt.$$

This consequence of the Cauchy-Schwartz inequality is entrusted to the reader as an exercise.

We now complete the proof of the theorem. We proceed to pick $\delta > 0$ sufficiently small so that $P(t) > \delta$ on $[a,b]$, and (calling upon a known fact in differential equations), also so that the following differential equation admits a solution w_δ:

$$w_\delta{}'(t) = \frac{w_\delta(t)^2}{P(t)-\delta} - Q(t),\ t \in [a,b].$$

(The hypothesized existence of the solution $w = w_0$ for $\delta = 0$ is crucial here for being able to assert that the solution w_δ of the perturbed equation exists for suitably small δ; see [28].) We then pick any variation y satisfying

$$\|y\| + \|y'\| \leqslant \min\ \big(\delta/4M, 1\big).$$

We then find (exactly as in Legendre's argument)

$$\int_a^b \left[(P-\delta)y'^2 + Qy^2\right] dt = \int_a^b (P-\delta)\big\{y' + (w_\delta y)/(P-\delta)\big\}^2\,dt \geqslant 0.$$

Applying the lemma, we deduce

$$\begin{aligned} J(x_*+y) - J(x_*) &\geqslant \tfrac{1}{2}\int_a^b \left[Py'^2 + Qy^2\right] dt - \tfrac{\delta}{4}\int_a^b y'^2\,dt \\ &= \tfrac{1}{2}\int_a^b \left[(P-\delta)y'^2 + Qy^2\right] dt + \tfrac{\delta}{4}\int_a^b y'^2\,dt \geqslant 0. \end{aligned}$$ □

14.9 Example. The function $x_* \equiv 0$ is an admissible extremal for the problem

$$\min \int_0^1 \left\{ \tfrac{1}{2}x'(t)^2 + t\,x'(t)\sin x(t)\right\} dt \,:\, x \in C^2[0,1],\ x(0) = x(1) = 0,$$

as the reader may verify. We wish to show that it provides a weak local minimum. We calculate

$$P(t) = 1, \quad Q(t) = -1.$$

Thus, the strengthened Legendre condition holds, and it suffices (by Theorem 14.8) to exhibit a solution w on $[0,1]$ of the differential equation $w' = w^2 + 1$. The function $w(t) = \tan t$ serves the purpose. Note that this w would fail to be defined, however, if the underlying interval were $[0,2]$ (say). □

The following exercise demonstrates that merely *pointwise* conditions such as the Euler equation and the strengthened Legendre condition cannot always imply optimality; some extra element must be introduced which refers to the interval $[a,b]$

or, more precisely, its length. We arrive in this way at a new idea: extremals can be optimal *in the short term*, without being optimal globally on their entire interval of definition.

14.10 Exercise. (Wirtinger's inequality) We study the putative inequality

$$\int_0^T x^2(t)\,dt \leqslant \int_0^T x'(t)^2\,dt$$

for smooth functions $x:[0,T]\to\mathbb{R}$ which vanish at 0 and T (here, $T>0$ is fixed). We rephrase the issue as follows: we study whether the function $x_*\equiv 0$ solves the following special case of the problem (P):

$$\min J(x) = \int_0^T \left[x'^2 - x^2\right] dt \;:\; \text{subject to } x\in C^2[a,b],\; x(0)=x(T)=0. \quad (*)$$

(a) Show that, whatever the value of T, the function x_* is an extremal and satisfies the strengthened Legendre condition on $[0,T]$.

(b) For any x admissible for $(*)$, we have $J(\lambda x)=\lambda^2 J(x)\ \ \forall\lambda>0$. Deduce from this homogeneity property that x_* is a weak local minimizer for $(*)$ if and only if it is a global minimizer.

(c) Let $T\leqslant 1$, and let x be admissible for $(*)$. Use the Cauchy-Schwarz inequality to prove

$$|x(t)|^2 \leqslant \int_0^T x'(s)^2\,ds,\;\; t\in[0,T].$$

Deduce that $J(x)\geqslant 0 = J(x_*)$, so that x_* does solve problem $(*)$.

(d) Now let $T\geqslant 4$. Show that the function x defined by $x(0)=0$ and

$$x'(t) = \begin{cases} 1 & \text{if } 0\leqslant t\leqslant 1 \\ 0 & \text{if } 1<t<T-1 \\ -1 & \text{if } T-1\leqslant t\leqslant T \end{cases}$$

satisfies $J(x)<0$. Note that x is Lipschitz, but not C^2; however, approximation (see Exer. 21.13) leads to the conclusion that for $T\geqslant 4$, the extremal x_* *fails* to solve $(*)$.

It follows that the optimality of x_* for $(*)$ *ceases* to hold at some value of T between 1 and 4. We surmise that it must be a notable number; we shall identify it in due course. □

Conjugate points. We know from local existence theorems that the differential equation that appears in the statement of Theorem 14.8 admits a solution w on an interval $[a,a+\varepsilon]$, for some $\varepsilon>0$ (in the presence of the other hypotheses). It follows that the extremal x_*, restricted to $[a,a+\varepsilon]$, is a weak local minimizer for

the truncated version of the basic problem for which it is admissible (that is, the problem whose boundary values at a and $a+\varepsilon$ are those of x_*). The difficulty is that $a+\varepsilon$ may have to be strictly less than b.

A half-century after Legendre, Jacobi found a way to calibrate the extent of an extremal's optimality. We now examine this theory, which is based upon a certain second-order differential equation. Let x_* be an extremal on $[a,b]$, and let P and Q be defined as before:

$$\begin{aligned} P(t) &= \Lambda_{vv}\big(t, x_*(t), x_*'(t)\big) \\ Q(t) &= \Lambda_{xx}\big(t, x_*(t), x_*'(t)\big) - \frac{d}{dt}\Lambda_{xv}\big(t, x_*(t), x_*'(t)\big). \end{aligned}$$

The **Jacobi equation** corresponding to x_* is the following second-order differential equation:

$$-\frac{d}{dt}\big\{P(t)\,u'(t)\big\} + Q(t)\,u(t) = 0, \;\; u \in C^2[a,b].$$

The somewhat unusual way in which this differential equation is expressed (as in a classical Sturm-Liouville problem) is traditional.

14.11 Definition. *The point τ in $(a,b]$ is said to be conjugate to a (relative to the given extremal x_*) if there is a nontrivial solution u of the associated Jacobi equation which satisfies $u(a) = u(\tau) = 0$.*

In seeking conjugate points, it turns out that *any* nontrivial solution u of Jacobi's equation vanishing at a can be used: any other such u will generate the *same* conjugate points (if any). Let us see why this is so. Consider two such functions u_1 and u_2. We claim that $u_1'(a) \neq 0$. The reason for this is that the only solution u of Jacobi's equation (a linear homogeneous differential equation of order two) satisfying $u(a) = u'(a) = 0$ is the zero function (by the well-known uniqueness theorem for the initial-value problem). Similarly, we have $u_2'(a) \neq 0$. It follows that for certain nonzero constants c, d, we have

$$c\,u_1'(a) + d\,u_2'(a) = 0.$$

But then the function $u := c\,u_1 + d\,u_2$ is a solution of Jacobi's equation which vanishes, together with its derivative, at a. Thus $u \equiv 0$; that is, u_2 is a nonzero multiple of u_1. It follows, then, that u_1 and u_2 have the same zeros, and hence determine the same conjugate points.

A nontrivial solution u of Jacobi's equation that vanishes at a has a *first* zero $\tau > a$, if it has one at all (for otherwise we find $u'(a) = 0$, a contradiction. Thus, it makes sense to speak of the *nearest* conjugate point τ (if any), which is located at a strictly positive distance to the right of a.

In the study of conjugate points, it is always assumed that the underlying extremal x_* satisfies the strengthened Legendre condition: $P(t) > 0 \;\; \forall t \in [a,b]$.

14.12 Theorem. (Jacobi 1838) *Let $x_* \in C^2[a,b]$ be an extremal of the basic problem* (P) *which satisfies the boundary conditions of* (P) *as well as the strengthened Legendre condition. Then*

(a) (Necessary condition) *If x_* is a weak local minimizer for* (P)*, there is no conjugate point to a in the interval (a,b).*

(b) (Sufficient condition) *Conversely, if there is no point conjugate to a in the interval $(a,b]$, then x_* is a weak local minimizer for* (P).

Proof. The proof of necessity is postponed (see Ex. 15.6). We prove here the sufficiency. Accordingly, let x_* be an admissible extremal satisfying the strengthened Legendre condition, and admitting no conjugate point in $(a,b]$.

Lemma. *Jacobi's equation*

$$-\frac{d}{dt}\{P(t)\,u'(t)\} + Q(t)\,u(t) = 0$$

admits a solution $\bar{u}$ on $[a,b]$ which is nonvanishing.

To see this, let us consider first the solution u_0 on $[a,b]$ of Jacobi's equation with initial condition $u(a) = 0,\ u'(a) = 1$ (such a solution exists because of the linearity of Jacobi's equation).

Since there is no conjugate point in the interval $(a,b]$ (by hypothesis), it follows that u_0 is nonvanishing on the interval $(a,b]$. Because u_0' is continuous, we can therefore find $\varepsilon > 0$ and $d \in (a,b)$ such that

$$u_0'(t) > \varepsilon\ (t \in [a,d]),\ \ |u_0(t)| > \varepsilon,\ \ (t \in [d,b]).$$

Now consider the solution u_η of the Jacobi equation satisfying the initial conditions $u(a) = \eta,\ u'(a) = 1$, where η is a small positive parameter. According to the "imbedding theorem" (see [28]) whereby solutions of differential equations depend continuously upon initial conditions, for η sufficiently small we have

$$\left|u_\eta{}'(t) - u_0'(t)\right| < \frac{\varepsilon}{2},\ \ |u_\eta(t) - u_0(t)| < \frac{\varepsilon}{2},\ \ t \in [a,b].$$

(In order to apply the imbedding theorem, we rewrite the second-order differential equation as a system of two first-order equations, in the usual way; the positivity of P is used in this.) Note that u_η clearly cannot vanish on $[d,b]$; it also cannot vanish on $[a,d]$, since we have $u_\eta(a) > 0$ and $u_\eta{}' > 0$ on $[a,d]$. Thus u_η is nonvanishing on $[a,b]$. This proves the lemma: take $\bar{u} = u_\eta$.

We now complete the proof of the theorem. We set

$$w(t) = -\bar{u}'(t)P(t)/\bar{u}(t),$$

which is possible because $\bar{u}$ is nonvanishing. This clever change of variables was discovered by Jacobi; it linearizes the differential equation that appears in the statement of Theorem 14.8, in the sense that, as follows by routine calculation, the resulting function w satisfies

$$w'(t) = \frac{w(t)^2}{P(t)} - Q(t)\,,\ t \in [a,b].$$

Thus x_* is revealed to be a weak local minimizer, by Theorem 14.8. □

14.13 Corollary. *Let $x_* \in C^2[a,b]$ be an extremal of the basic problem* (P) *which satisfies the boundary conditions of* (P) *as well as the strengthened Legendre condition. Suppose there exists a solution u of Jacobi's equation which does not vanish on $[a,b]$. Then x_* is a weak local minimizer.*

Proof. As shown in the proof of Theorem 14.12, u induces a solution of the differential equation that appears in the statement of Theorem 14.8; thus, the conclusion follows from that result. □

14.14 Example. Consider the basic problem

$$\min \int_0^1 x'(t)^3\,dt \;:\; x \in C^2[0,1],\ x(0) = 0,\ x(1) = 1.$$

The Euler equation is

$$\frac{d}{dt}\left\{3x'^2\right\} = 0,$$

which implies that x' is constant. Thus, the unique admissible extremal is $x_*(t) = t$. The corresponding functions P and Q are given by $P(t) = 6$ and $Q = 0$. It follows that the strengthened Legendre condition holds along x_*, and the Jacobi equation is

$$-\frac{d}{dt}\left\{6u'\right\} = 0.$$

A suitable solution of this equation (for finding conjugate points) is $u(t) = t$. Since this has no zeros in $(0,1]$, and hence generates no conjugate points, we deduce from Theorem 14.12 that x_* provides a weak local minimum. □

14.15 Exercise. Prove that the problem considered in Example 14.14 admits no weak local maximizer. □

In the classical situations of mechanics, extremals of the action functional satisfy the strengthened Legendre condition, because the kinetic energy term is positive definite and quadratic. Jacobi's theorem therefore confirms the principle of least action as a true minimization assertion: the action is in fact minimized *locally* and *in the short term* by any extremal (relative to the points that it joins).

14.16 Exercise. Consider the soap bubble problem (Example 14.5), with

$$[a,b] = [0,T],\ (T>0),\ A = 1,\ B = \cosh T.$$

Let x_* be the catenary $\cosh t$. Prove that, if T is sufficiently small, then x_* provides a weak local minimum for the problem. □

14.17 Exercise. (Wirtinger's inequality, continuation) Find Jacobi's equation for the extremal and Lagrangian of the problem arising from the Wirtinger inequality (Exer. 14.10). Show that the point $\tau = \pi$ (a notable number) is the first point conjugate to $a = 0$. Deduce the following result:[4]

Proposition. *Let $T \in [0,\pi]$. Then, for any $x \in C^2[0,T]$ which vanishes at 0 and T, we have*

$$\int_0^T x(t)^2\,dt \leqslant \int_0^T x'(t)^2\,dt.$$

The general inequality fails if $T > \pi$. □

14.18 Exercise. We study the following problem in the calculus of variations:

$$\text{minimize} \int_0^T e^{-\delta t}\big(x'(t)^2 - x(t)^2\big)\,dt \,:\, x \in C^2[0,T],\ x(0)=0,\ x(T)=0,$$

where $T > 0$ and $\delta \geqslant 0$ are given.

(a) Show that $x_* \equiv 0$ is a weak local minimizer when $\delta \geqslant 2$, for any value of $T > 0$.

(b) If $\delta \geqslant 2$, prove that x_* is in fact a global minimizer. [Hint: $J(\lambda x) = \lambda^2 J(x)$.]

(c) When $\delta < 2$, show that, for a certain $\tau > 0$, x_* is a local minimizer if $T < \tau$, but fails to provide a local minimum if $T > \tau$. □

14.3 Two variants of the basic problem

We conclude this chapter's survey of the classical theory with two well-known variants of the underlying problem.

The transversality condition. In certain variational problems, the endpoint values of the competing functions x are not fully prescribed. In such cases, the extra flexibility at the boundary gives rise to additional conclusions in the necessary conditions, conclusions that say something about the initial and/or final values of x. These are known as *transversality conditions*.

[4] See Exer. 21.14 for an equivalent version of Wirtinger's inequality, one that is formulated on intervals of arbitrary length.

The following provides a simple example. We consider the problem of minimizing

$$\ell\big(x(b)\big) + \int_a^b \Lambda\big(t, x(t), x'(t)\big)\, dt$$

over the functions $x \in C^2[a,b]$ satisfying the initial condition $x(a) = A$. The given function ℓ (which we take to be continuously differentiable) corresponds to an extra cost term that depends on the (now unprescribed) value of $x(b)$.

14.19 Theorem. *Let x_* be a weak local minimizer of the problem. Then x_* is an extremal for Λ, and x_* also satisfies the following* **transversality condition:**

$$-\Lambda_v\big(b, x_*(b), x_*'(b)\big) = \ell'\big(x_*(b)\big).$$

The main point to be retained is that the extra information provided by the transversality condition exactly compensates for the fact that $x_*(b)$ is now unknown. Thus the overall balance between known and unknown quantities is preserved: there is *conservation of information*. We shall see other instances later of this general principle for necessary conditions.

Proof. It is clear that x_* is a weak local minimizer for the version of the original problem (P) in which we impose the final constraint corresponding to $B := x_*(b)$. Thus, x_* is an extremal by Theorem 14.2.

Let us now choose any function $y \in C^2[a,b]$ for which $y(a) = 0$ (but leaving $y(b)$ unspecified). We define g as follows:

$$g(\lambda) = \ell\big(x_*(b) + \lambda y(b)\big) + J(x_* + \lambda y).$$

It follows that g has a local minimum at $\lambda = 0$; thus $g'(0) = 0$. As in the proof of Theorem 14.2, this leads to

$$\ell'\big(x_*(b)\big)\, y(b) + \int_a^b \big[\alpha(t)\, y(t) + \beta(t)\, y'(t)\big]\, dt = 0.$$

Since $\alpha = \beta'$ (as we know from the Euler equation), integration by parts shows that the integral is equal to $\beta(b)\, y(b)$. We derive therefore

$$\big[\ell'\big(x_*(b)\big) + \beta(b)\big]\, y(b) = 0.$$

Since $y(b)$ is arbitrary, we deduce $\ell'\big(x_*(b)\big) + \beta(b) = 0$, which is the desired conclusion. □

14.20 Exercise. Given that the following problem has a unique solution x_*:

$$\text{minimize} \int_0^3 \big(\tfrac{1}{2} x'(t)^2 + x(t)\big)\, dt \; : \; x \in C^2[0,3],\; x(0) = 0,$$

show that x_* is of the form $t^2/2+ct$. Use the transversality condition to show that $c=-3$. Find the solution when the problem is modified as follows:

$$\text{minimize } x(3)+\int_0^3 \big(\tfrac{1}{2}x'(t)^2+x(t)\big)\,dt \;:\; x\in C^2[0,3],\; x(0)=0. \qquad \square$$

The isoperimetric problem. This phrase refers to the classical problem of minimizing the same functional $J(x)$ as in the basic problem (P), under the same boundary conditions, but under an additional equality constraint defined by a functional of the same type as J:

$$\int_a^b \psi\big(t,x(t),x'(t)\big)\,dt = 0.$$

The method of multipliers was introduced in this context by Euler (yes, in that very same monograph), but, as we know, is most often named after Lagrange, who made systematic use of it in his famous treatise on mechanics. Because of our experience with the multiplier rule in optimization, the reader will find that the next result has a familiar look.

14.21 Theorem. *Let $x_*\in C^2[a,b]$ be a weak local minimizer for the isoperimetric problem, where Λ and ψ are C^2. Then there exists $(\eta,\lambda)\neq 0$, with $\eta=0$ or 1, such that x_* is an extremal for the Lagrangian $\eta\Lambda+\lambda\psi$.*

Proof. We merely sketch the proof; a much more general multiplier rule will be established later. The idea is to derive the conclusion from Theorem 9.1, for the purposes of which we define

$$f(x)=J_\Lambda(x_*+x),\quad h(x)=J_\psi(x_*+x),$$

where J_Λ and J_ψ are the integral functionals corresponding to Λ and ψ, and where x lies in the vector space $X=C_0^2[a,b]$ consisting of those elements in $C^2[a,b]$ which vanish at a and b. We may conveniently norm X by $\|x\|_X=\|x''\|_\infty$, turning it into a Banach space.

Then $f(x)$ attains a local minimum in X relative to $h(x)=0$ at $x=0$. It is elementary to show that f and h are continuously differentiable. It follows from Theorem 9.1 that for (η,λ) as described, we have $(\eta f+\lambda h)'(0;y)=0$ for every $y\in X$. As in the proof of Theorem 14.2, this implies that x_* is an extremal of $\eta\Lambda+\lambda\psi$. $\square$

14.22 Exercise. A homogeneous chain of length L is attached to two points (a,A) and (b,B), where $a<b$. Hanging in equilibrium, it describes a curve $x(t)$ which minimizes the potential energy, which can be shown to be of the form

$$\sigma\int_a^b x(t)\sqrt{1+x'(t)^2}\,dt,$$

where $\sigma > 0$ is the (constant) mass density. Thus, we seek to minimize this functional relative to all curves in $C^2[a,b]$ having the same endpoints and the same length:

$$x(a) = A\,,\ x(b) = B\,,\ \int_a^b \sqrt{1+x'(t)^2}\,dt \ = \ L\,.$$

We assume that the chain is long enough to make the problem meaningful: L is greater than the distance between (a,A) and (b,B).

Show that if x_* is a solution to the problem, then, for some constant λ, the function $x_*+\lambda$ is an extremal of the Lagrangian $x\sqrt{1+v^2}$. Invoke Example 14.5 to reveal that x_* is a translate of a catenary. □

We remark that the problem treated in the exercise above is a continuous version of the discrete one considered in Exer. 13.5. The conclusion explains the etymology: "catena" means "chain" in Latin.

14.23 Exercise. Assuming that a solution exists (this will be confirmed later), solve the following isoperimetric problem:

$$\min \int_0^\pi x'(t)^2\,dt \,:\, x \in C^2[0,\pi]\,,\ \int_0^\pi x(t)^2\,dt \ = \pi/2\,,\ x(0) = x(\pi) = 0.$$

(The analysis is continued in Exer. 16.12.) □

14.24 Exercise. By examining the necessary conditions for the problem

$$\min \int_0^\pi x(t)^2\,dt \,:\, x \in C^2[0,\pi]\,,\ \int_0^\pi x'(t)^2\,dt \ = \pi/2\,,\ x(0) = x(\pi) = 0,$$

show that it does *not* admit a solution. What is the infimum in the problem? □

Chapter 15

Nonsmooth extremals

Is it always satisfactory to consider only smooth solutions of the basic problem, as we have done in the previous chapter? By the middle of the 19th century, the need to go beyond continuously differentiable functions had become increasingly apparent.[1]

15.1 Example. Consider the following instance of the basic problem:

$$\text{minimize } J(x) = \int_{-1}^{1} x(t)^2 \left[x'(t) - 1\right]^2 dt \text{ subject to } x(-1) = 0,\ x(1) = 1.$$

Note that $J(x) \geqslant 0\ \ \forall x$. Let x lie in $C^1[-1,1]$ and satisfy the given boundary conditions. Then there exists $\tau \in (-1,1)$ such that $x'(\tau) = 1/2$. Therefore there is a neighborhood of τ in which x vanishes at most once and $x' \neq 1$; it follows that $J(x) > 0$. Consider now the continuous, piecewise-smooth function

$$x_*(t) = \begin{cases} 0 & \text{if } -1 \leqslant t \leqslant 0 \\ t & \text{if } 0 \leqslant t \leqslant 1, \end{cases}$$

which has a *corner* at $t = 0$. Then $J(x_*) = 0$. Furthermore, it is easy to show that the infimum of J over the admissible functions in $C^1[-1,1]$ (or in $C^2[-1,1]$) is 0, an infimum that is therefore not attained in that class. To summarize, the problem has a natural solution in the class of piecewise-smooth functions, but has no solution in that of smooth functions. □

The example shows that the very existence of solutions is an impetus for admitting nonsmooth arcs. The need also became apparent in physical applications (soap bubbles, for example, generally have corners and creases). Spurred by these considerations, the theory of the basic problem was extended to the context of piecewise-smooth functions. In this chapter, we develop this theory, but within the more general class of *Lipschitz* functions x; that is, absolutely continuous functions whose

---

[1] The calculus of variations seems to have been the first subject to acknowledge a need to consider nonsmooth functions.

F. Clarke, *Functional Analysis, Calculus of Variations and Optimal Control*, Graduate Texts in Mathematics 264, DOI 10.1007/978-1-4471-4820-3_15,

derivatives x' exist almost everywhere and are essentially bounded. Of course, we are only able to deal with this class since we have the benefit of Lebesgue's legacy in measure and integration (which was not available until roughly 1900).

Hypotheses. We extend the setting of the basic problem in one more way, by allowing x to be vector-valued. Thus, the admissible class now becomes the Lipschitz functions mapping $[a,b]$ to $\mathbb{R}^n$, for which we simply write $\text{Lip}[a,b]$. Each component of x, then, is an element of the space $\text{AC}^\infty[a,b]$ introduced in Example 1.13. The phrase "basic problem" refers in this chapter to

$$\text{minimize } J(x) : x \in \text{Lip}[a,b],\ x(a) = A,\ x(b) = B. \qquad \textbf{(P)}$$

The hypotheses on $\Lambda(t,x,v)$ are also weakened: we suppose that the gradients Λ_x and Λ_v exist and, together with Λ, are continuous functions of (t,x,v). Note that under these hypotheses, the functional

$$J(x) = \int_a^b \Lambda\big(t, x(t), x'(t)\big)\,dt$$

is well defined (as a Lebesgue integral) for every $x \in \text{Lip}[a,b]$.

It can be shown (see Exer. 21.13) that a minimum over $C^2[a,b]$ is also a minimum over $\text{Lip}[a,b]$, so the extension of the classical theory to Lipschitz functions is a faithful one.

15.1 The integral Euler equation

The first order of business is to see what becomes of the basic necessary condition, namely the Euler equation, when the basic problem (P) is posed over the class $\text{Lip}[a,b]$.

The definition of weak local minimum x_* is essentially unchanged: the minimum is relative to the admissible functions x that satisfy $\|x - x_*\| \leqslant \varepsilon$ and $\|x' - x_*'\| \leqslant \varepsilon$ (recall that $\|\cdot\|$ refers to the L^∞ norm). However, we must inform the reader of the regrettable fact that the proof of Theorem 14.2 *fails* when $x \in \text{Lip}[a,b]$. The problem is that the function $t \mapsto \Lambda_v(t,\, x_*(t),\, x_*'(t))$ is no longer differentiable, so the integration by parts that was used in the argument can no longer be justified. There is, however, a good way to extend the Euler equation to our new situation.

15.2 Theorem. (du Bois-Raymond 1879) *Let $x_* \in Lip[a,b]$ be a weak local minimizer for the basic problem* (P). *Then x_* satisfies the* **integral Euler equation:** *for some constant $c \in \mathbb{R}^n$, we have*

$$\Lambda_v\big(t, x_*(t), x_*'(t)\big) = c + \int_a^t \Lambda_x\big(s, x_*(s), x_*'(s)\big)\,ds,\ \ t \in [a,b] \text{ a.e.}$$

Proof. We define $g(\lambda)$ as we did in the proof of Theorem 14.2, where now the variation y lies in $\mathrm{Lip}_0[a,b]$, the set of Lipschitz functions on $[a,b]$ that vanish at a and b. The difference quotient whose limit is $g'(0)$ is given by

$$\int_a^b \frac{\Lambda(t,x_*+\lambda y,x_*'+\lambda y')-\Lambda(t,x_*,x_*')}{\lambda}\,dt.$$

Our hypotheses imply that Λ is Lipschitz with respect to (x,v) on bounded sets. Since all the functions appearing inside the integral above are bounded, there is a constant K such that, for all λ near 0, the integrand is bounded by $K|(y(t),y'(t))|$. It now follows from Lebesgue's dominated convergence theorem that $g'(0)$ exists and is given by

$$g'(0) = \int_a^b \big[\alpha(t)\bullet y(t)+\beta(t)\bullet y'(t)\big]\,dt = 0,$$

where (as before, but now with gradients instead of partial derivatives)

$$\alpha(t) = \Lambda_x\big(t,x_*(t),x_*'(t)\big),\quad \beta(t) = \Lambda_v\big(t,x_*(t),x_*'(t)\big).$$

Note that these are essentially bounded functions. We apply integration by parts, but now to the *first* term in the integral; this yields

$$\int_a^b \left[\beta(t)-\int_a^t \alpha(s)\,ds\right]\bullet y'(t)\,dt = 0.$$

For any $c\in\mathbb{R}^n$, we also have

$$\int_a^b \left[\beta(t)-c-\int_a^t \alpha(s)\,ds\right]\bullet y'(t)\,dt = 0.$$

This holds, then, for all Lipschitz variations y and all constants $c\in\mathbb{R}^n$. We now choose c such that the function

$$y(t) = \int_a^t \left[\beta(t)-c-\int_a^t \alpha(s)\,ds\right]dt$$

defines a variation (that is, such that $y(b)=0$). With this choice of y, we discover

$$\int_a^b \left|\beta(t)-c-\int_a^t \alpha(s)\,ds\right|^2 dt = 0,$$

which implies the desired conclusion. □

When $x_*\in C^2[a,b]$ and $n=1$, it is evident that the integral Euler equation is equivalent to the original one obtained in Theorem 14.2 (which can be thought of as the "differentiated form"). So we have lost nothing in this new formulation. Note, however, that when $n>1$, we are dealing with a *system* of equations, since the gradients Λ_x and Λ_v are vector-valued. In other words, there are n unknown functions involved, the n component functions of x.

15.3 Example. Consider a particle of mass m that is free to move in the x-y plane under the sole influence of a conservative force field F given by a potential V: $F(x,y) = -\nabla V(x,y)$. The action in this case is the integral functional

$$\int_{t_0}^{t_1} \Big\{ \frac{m}{2} \big|\big(x'(t), y'(t)\big)\big|^2 - V\big(x(t), y(t)\big) \Big\}\, dt.$$

Thus, the Lagrangian is given by

$$\Lambda(t,x,y,v,w) = \frac{m}{2}\,|(v,w)|^2 - V(x,y).$$

In its original (differentiated) form, the Euler equation, which is now a system of two differential equations, asserts

$$\frac{d}{dt}\big[m x'(t), m y'(t)\big] = -\nabla V\big(x(t), y(t)\big) = F\big(x(t), y(t)\big),$$

which we recognize as Newton's Law: $F = mA$. We ask the reader to verify that the integral form of the Euler equation gives exactly the same conclusion in this case. □

The costate. In the context of Theorem 15.2, let us define

$$p(t) = c + \int_a^t \Lambda_x\big(s, x_*(s), x_*'(s)\big)\, ds.$$

Then the function $p : [a,b] \to \mathbb{R}^n$ belongs to $\mathrm{Lip}[a,b]$ and satisfies

$$\big(p'(t), p(t)\big) = \nabla_{x,v}\Lambda\big(t, x_*(t), x_*'(t)\big) \text{ a.e.}$$

This is a convenient way to write the integral Euler equation of Theorem 15.2, one that highlights the fact that the mapping

$$t \mapsto \Lambda_v\big(t, x_*(t), x_*'(t)\big)$$

must have removable discontinuities (even though x_*' is not necessarily continuous).[2] In some cases, as we shall see, this implies the continuity of x_*'.

In classical mechanics, p is referred to as the *generalized momentum*; another synonym is *adjoint variable*. In optimal control, where p figures in the celebrated Pontryagin maximum principle, p is often called the *costate*; we shall retain this terminology.

In addition to its role in the integral Euler equation, the costate is also convenient for expressing the appropriate transversality condition when the values of x at the

[2] This is known classically as the *first* Erdmann condition, whereas the conclusion of Prop. 14.4 is the *second* Erdmann condition. There does not seem to be a third.

boundary are not fully prescribed. In the context of Theorem 14.19, the condition we found, extended now to $n > 1$, was

$$-p(b) = \nabla \ell\big(x_*(b)\big).$$

Later on, we shall admit boundary costs and constraints of a more general nature, and corresponding transversality conditions. If, for example, the endpoint constraint is given by $x(b) \in E$, then the resulting transversality condition is

$$-p(b) \in N_E^L\big(x(b)\big).$$

Facts such as these suggest the possibility of using (x, p) as new coordinates, a procedure which Legendre was the first to employ. In classical mechanics, the *phase coordinates* (x, p) can be used to rewrite the integral Euler equation in *Hamiltonian form* as follows:

$$-p'(t) = H_x\big(t, x(t), p(t)\big), \quad x'(t) = H_p\big(t, x(t), p(t)\big),$$

where the Hamiltonian function H is obtained from the Lagrangian Λ via the *Legendre transform* (the precursor of the Fenchel transform in convex analysis). The transform is explained in Exer. 21.28; it continues to play an important role in mathematical physics to this day.

In general, the Euler equation may admit more solutions in $\mathrm{Lip}[a,b]$ than in $C^2[a,b]$, as illustrated by the following.

15.4 Example. Consider the basic problem

$$\min \int_0^1 x'(t)^3\, dt, \; x(0) = 0, \; x(1) = 1,$$

which we have met in Example 14.14, in the context of $C^2[0,1]$. The integral Euler equation is

$$\big(p'(t), p(t)\big) = \big(0, 3x'(t)^2\big) \text{ a.e.},$$

which implies $x'(t)^2 = c^2$, for some constant c. If x is restricted to $C^2[a,b]$ by assumption, then there results a unique admissible extremal: $x(t) = t$. If, however, we work in $\mathrm{Lip}[a,b]$, then there are infinitely many possibilities, including (for any $c > 1$) any piecewise affine admissible function whose derivative is almost everywhere $\pm c$. $\square$

The last example underscores the need for *additional* conditions that will help reduce the number of candidates. The Weierstrass necessary condition that we shall meet later plays a useful role in this vein, especially when it is extended to problems with additional constraints. Another important tool is that of *regularity theorems*, the next topic.

15.2 Regularity of Lipschitz solutions

A *regularity theorem* is one which affirms that, under some additional hypotheses, solutions of the basic problem, which initially are known to lie merely in the class in which (P) is posed (currently, $\mathrm{Lip}[a,b]$), actually lie in a *smaller* class of more regular functions (below, $C^1[a,b]$). This is a way of having your cake and eating it too: we allow nonsmooth functions in formulating the problem, yet we obtain solutions that are smooth.

15.5 Theorem. *Let $x_* \in Lip[a,b]$ satisfy the integral Euler equation, where, for almost every $t \in [a,b]$, the function $v \mapsto \Lambda(t,x_*(t),v)$ is strictly convex. Then x_* lies in $C^1[a,b]$.*

Proof. The goal is to find a continuous function $\bar{v}$ on $[a,b]$ which agrees with x_*' almost everywhere. For such a function, we have

$$x_*(t) = x_*(a) + \int_a^t x_*'(s)\,ds = x_*(a) + \int_a^t \bar{v}(s)\,ds \;\; \forall\, t \in [a,b],$$

from which it follows that $x_* \in C^1[a,b]$. We define the set

$$W = \big\{ t \in (a,b) : x_*'(t) \text{ exists and } p(t) = \Lambda_v\big(t, x_*(t), x_*'(t)\big)\big\}.$$

Then W is of full measure. Fix any $\tau \in [a,b]$, and let $\{s_i\}$ and $\{t_i\}$ be any two sequences in W converging to τ, and such that

$$\ell_s := \lim_{i\to\infty} x_*'(s_i), \quad \ell_t := \lim_{i\to\infty} x_*'(t_i)$$

both exist. Passing to the limit in the equation $p(s_i) = \Lambda_v(s_i, x_*(s_i), x_*'(s_i))$ yields $p(\tau) = \Lambda_v(\tau, x_*(\tau), \ell_s)$. Similarly, we derive $p(\tau) = \Lambda_v(\tau, x_*(\tau), \ell_t)$. It follows that the strictly convex function $v \mapsto \Lambda(\tau, x_*(\tau), v)$ has the same gradient at ℓ_s and ℓ_t, whence $\ell_s = \ell_t$ (see Exer. 4.17).

Let us now define a function $\bar{v}$ on $[a,b]$ as follows: take any sequence $\{s_i\}$ in W converging to τ and such that $\lim_i x_*'(s_i)$ exists (such sequences exist because W is of full measure, and because x_*' is bounded); set $\bar{v}(\tau) = \lim_i x_*'(s_i)$. In view of the preceding remarks, $\bar{v}$ is well defined. But $\bar{v}$ is continuous by construction, as is easily seen, and agrees with x_*' on W. □

Note that the theorem applies to problems involving the classical action (such as Example 15.3), but fails to apply to that of Example 15.1, for instance.

15.6 Exercise. As an application of Theorem 15.5, we prove the necessity of Jacobi's condition; the setting is that of Theorem 14.12. We reason *ad absurdum*, by supposing that a conjugate point $\tau \in (a,b)$ exists.

Recall that the proof of Theorem 14.7 showed that, for every $y \in \mathrm{Lip}_0[a,b]$, we have

$$I(y) := \int_a^b \left\{ P(t)\, y'(t)^2 + Q(t)\, y(t)^2 \right\} dt \;\geqslant 0.$$

(a) By assumption there is a nontrivial solution u of Jacobi's equation vanishing at a and τ. Prove that $u'(\tau) \neq 0$, and that

$$\int_a^\tau \left\{ P(t)\, u'(t)^2 + Q(t)\, u(t)^2 \right\} dt \;= 0.$$

(b) Extend u to $[a,b]$ by setting it equal to 0 between τ and b. It follows that u minimizes the functional $I(y)$ above relative to the arcs $y \in \mathrm{Lip}_0[a,b]$. Apply Theorem 15.5 to obtain a contradiction. □

Higher regularity. It is possible to go further in the regularity theory, and show that, under suitable conditions, the solutions of the basic problem inherit the full regularity of the Lagrangian.

15.7 Theorem. (Hilbert-Weierstrass c. 1875) *Let $x_* \in Lip[a,b]$ satisfy the integral Euler equation, where Λ is of class C^m $(m \geqslant 2)$ and satisfies*

$$t \in [a,b], v \in \mathbb{R}^n \implies \Lambda_{vv}(t, x_*(t), v) > 0 \text{ (positive definite)}.$$

Then x_* belongs to $C^m[a,b]$.

Proof. The hypotheses imply the strict convexity of the function $v \mapsto \Lambda(t, x_*(t), v)$ for each t. It follows from Theorem 15.5 that x_* belongs to $C^1[a,b]$. We deduce that the costate $p(\cdot)$ belongs to $C^1[a,b]$, since, by the integral Euler equation,

$$p'(t) = \Lambda_x\big(t, x_*(t), x_*'(t)\big) \text{ (continuous on } [a,b]).$$

Note that for fixed t, the unique solution v of the equation $p(t) = \Lambda_v(t, x_*(t), v)$ is $v = x_*'(t)$ (since $\Lambda_{vv} > 0$). It follows from the implicit function theorem that x_*' lies in $C^1[a,b]$, since p is C^1 and Λ_v is C^{m-1} with $m \geqslant 2$. Thus x_* lies in $C^2[a,b]$. This conclusion now implies $p' \in C^1[a,b]$ and $p(\cdot) \in C^2[a,b]$. If $m > 2$, we may iterate the argument given above to deduce $x_*' \in C^2[a,b]$. We continue in this fashion until we arrive at $x_*' \in C^{m-1}[a,b]$; that is, $x_* \in C^m[a,b]$. □

15.8 Exercise. We consider the basic problem (P) when Λ has the form

$$\Lambda(t,x,v) = g(x)\sqrt{1 + |x'|^2}\,,$$

where g is C^m $(m \geqslant 2)$. Let $x_* \in \mathrm{Lip}[a,b]$ be a weak local minimizer, and suppose that $g(x_*(t)) > 0 \;\forall t \in [a,b]$. Prove that $x_* \in C^m[a,b]$. □

15.3 Sufficiency by convexity

The following inductive method *par excellence*, a basic fact concerning convex Lagrangians, went completely unobserved in the classical theory.[3]

15.9 Theorem. *Let* $x_* \in Lip[a,b]$ *be admissible for* (P) *and satisfy the integral Euler equation. Suppose that* $\Lambda(t,x,v)$ *is convex in* (x,v) *for each* t. *Then* x_* *is a global minimizer for* (P).

Proof. Let $x \in \mathrm{Lip}[a,b]$ be any admissible function for (P), and let p be the costate corresponding to x_* in the integral Euler equation. Then

$$\begin{aligned} J(x) - J(x_*) &= \int_a^b \bigl\{ \Lambda\bigl(t,x,x'\bigr) - \Lambda\bigl(t,x_*,x_*'\bigr) \bigr\}\,dt \\ &\geqslant \int_a^b \bigl(p',p\bigr) \bullet \bigl(x - x_*, x' - x_*'\bigr)\,dt \end{aligned}$$

(by the subgradient inequality, since $(p',p) = \nabla_{x,v}\Lambda(t,x_*,x_*')$ a.e.)

$$= \int_a^b (d/dt)\bigl\{ p \bullet (x - x_*) \bigr\}\,dt = 0,$$

since x and x_* agree at a and b. □

Remark. In Theorem 15.9, it is clear that if x_* happens to lie in $C^2[a,b]$, then it follows that x_* yields a global minimum relative to $C^2[a,b]$. The conclusion differs in several respects, however, from the classical sufficient conditions obtained by Legendre and Jacobi (Theorem 14.12). Aside from the reduced regularity hypotheses, there is no need for the strengthened Legendre condition, and (especially) one asserts the presence of a *global* minimum rather than a weak local one. In addition, it is easy to adapt the proof of Theorem 15.9 to certain cases in which the problem includes side constraints of the form (for example)

$$x(t) \in S, \quad x'(t) \in V,$$

where S and V are convex sets.

15.10 Example. Geodesics The problem of finding the curve of shortest length joining two distinct given points (the *geodesic*) is a venerable one in the subject. In the case of curves in the plane that join (a,A) and (b,B), the problem is naturally expressed in terms of two unknown functions $(x(t), y(t))$ parametrized on a given interval.

[3] This is but one example showing that our revered ancestors did not know about convex functions and their useful properties.

It is customary to take x and y to be Lipschitz, with $x'^2 + y'^2 \geqslant \varepsilon > 0$. Parametrizing by arc length would yield $x'^2 + y'^2 = 1$ a.e., but we prefer to fix $[0,1]$ as the underlying parameter interval (with no loss of generality). Then the functional to be minimized is

$$\int_0^1 \left\{ x'(t)^2 + y'(t)^2 \right\}^{1/2} dt\,,$$

subject to the given boundary conditions

$$\big(x(0), y(0)\big) = (a, A)\,, \quad \big(x(1), y(1)\big) = (b, B).$$

The integral Euler equation (a system of two differential equations) asserts the existence of constants (c,d) satisfying

$$c = \frac{x'}{\left\{x'^2 + y'^2\right\}^{1/2}}\,, \quad d = \frac{y'}{\left\{x'^2 + y'^2\right\}^{1/2}}\,.$$

This implies that x' and y' are constant, as the reader may show. Thus, the shortest curve is a line segment (a reassuring conclusion).

When $a < b$, it is reasonable to expect that the class of curves of interest may be restricted to those which can be parametrized in the form $(t, y(t))$, in which case we return to the case of a single unknown function. In general, however, several unknown functions will be involved, as in the general problem of finding geodesics on a given surface S in $\mathbb{R}^3$, where S is described by the equation $\psi(x,y,z) = 0$. Then the problem becomes that of minimizing

$$\int_0^1 \left\{ \left(\psi_x x'\right)^2 + \left(\psi_y y'\right)^2 + \left(\psi_z z'\right)^2 \right\}^{1/2} dt$$

subject to not only the boundary conditions, but also to a pointwise constraint

$$\psi\big(x(t), y(t), z(t)\big) = 0 \ \ \forall\, t \in [0,1]$$

that is imposed all along the admissible arcs. We shall discuss this type of problem, for which a multiplier rule exists, later in Chapter 17. Note the distinction with the isoperimetric problem: there are an *infinite* number of equality constraints here, one for each t.

In some cases the surface constraint can be made implicit by a suitable choice of coordinates, and we obtain a simpler form of the geodesic problem. Let us examine one such case now, that of the cylinder S in $\mathbb{R}^3$ defined by

$$\psi(x, y, z) = x^2 + y^2 - 1 = 0.$$

Then any curve on S may be parametrized in the form

$$\big(x(t), y(t), z(t)\big) = \big(\cos\theta(t), \sin\theta(t), z(t)\big), \ \ 0 \leqslant t \leqslant 1$$

for certain functions θ and z. And conversely, any curve parametrized this way necessarily lies on S.

Suppose (without loss of generality) that the initial point is given by $(\theta,z)=(0,0)$, and the final point by (θ_f,z_f), with $\theta_f \in (-\pi,\pi]$ and $(\theta_f,z_f) \neq (0,0)$. In this context we have

$$\big(\psi_x x'\big)^2+\big(\psi_y y'\big)^2+\big(\psi_z z'\big)^2=[(-\sin\theta)\theta']^2+[(\cos\theta)\theta']^2+z'^2=\theta'^2+z'^2,$$

so that the geodesic problem reduces to the following:

$$\min \int_0^1 \big\{\,\theta'(t)^2+z'(t)^2\big\}^{1/2} dt \text{ subject to}$$
$$(\theta,z)(0)=(0,0),\ (\theta,z)(1)=(\theta_f+2\pi k,z_f),\ k\in\mathbb{Z},$$

where the term $2\pi k$ reflects the possibility of winding around the cylinder several times. For fixed k, this is the same problem as that of finding geodesics in the (θ,z) plane, for which we know (see above) that the extremals are affine, whence:

$$\big(\theta(t),z(t)\big)=\big(t(\theta_f+2\pi k),t\,z_f\big),\ t\in[0,1].$$

The resulting cost is calculated to be

$$\big\{\big(\theta_f+2\pi k\big)^2+z_f^2\big\}^{1/2},$$

which is minimized relative to k by the choice $k=0$. Thus the geodesic describes a *helix* $(\theta(t),z(t))=(t\,\theta_f,t\,z_f)$ having at most a half turn around the z-axis.[4]

The analysis as it stands can quite properly be criticized on the basis that we do not know that a shortest curve actually exists. (Only a mathematician would worry about this; the reader should realize that we mean that as a compliment.) We may address this point, however, by exploiting convexity.

The Lagrangian here is given by

$$\Lambda(t,\theta,z,v,w)=\big\{v^2+w^2\big\}^{1/2}.$$

Since this function is convex in (θ,z,v,w), it follows from Theorem 15.9 that the extremal globally minimizes the integral (subject to the boundary conditions). □

15.11 Exercise.

(a) In the context of Theorem 15.9, prove that if the convexity of Λ with respect to (x,v) is *strict* for almost every t, then x_* is the unique solution of (P). Do we require strict convexity with respect to both variables?

[4] When $\theta_f=\pi$, the final point is antipodal to the initial point, and there is another (equally good) possibility: $k=-1$. When $\theta_f=0$, the geodesic reduces to a vertical segment.

(b) Prove that the admissible extremal found in Exer. 14.3 (c) is a unique global minimizer for the problem. □

Another variant of the theme whereby convexity renders necessary conditions sufficient is provided by the following.

15.12 Exercise. Consider the problem

$$\min \ \ell\big(x(b)\big)+\int_a^b \Lambda\big(t,x,x'\big)\,dt \ : \ x\in \operatorname{Lip}[a,b],\ x(a)=A,$$

where ℓ is convex and differentiable, and where Λ is convex in (x,v). Let x_* be admissible for the problem, and suppose that x_* satisfies the integral Euler equation for Λ, together with the transversality condition of Theorem 14.19 (now for $n \geqslant 1$):

$$-p(b) \,=\, \nabla\ell\big(x_*(b)\big).$$

Prove that x_* is a global minimizer for the problem. □

The following illustrates the use of convexity in connection with an isoperimetric problem.

15.13 Exercise. Consider the problem (Q):

$$\min \int_0^{\pi} x'(t)^2\,dt \ : \ x\in \operatorname{Lip}[0,\pi],\ \int_0^{\pi}(\sin t)\,x(t)\,dt = 1,\ x(0)=0,\ x(\pi)=\pi.$$

(a) Suppose for the moment that a solution x_* exists, and belongs to $C^2[0,\pi]$. Show that the necessary condition provided by Theorem 14.21 holds in normal form ($\eta=1$), and use it to identify x_*.

$$\Big[\text{Given: } \int_0^{\pi}\sin^2 t\,dt \,=\, \pi/2,\ \int_0^{\pi} t\sin t\,dt \,=\, \pi.\Big]$$

(b) Exploit the convexity in (x,v) of the Lagrangian $\Lambda_+ = v^2+\lambda(\sin t)x$ to deduce that the x_* identified above is in fact a global minimum for the problem (Q). □

15.4 The Weierstrass necessary condition

There remains one fundamental necessary condition in the classical theory that the reader has not yet met. It applies to a different, stronger type of local minimum than the weak local minimum we have considered so far.

A function x_* admissible for (P) is said to provide a **strong local minimum** for the problem if there exists $\varepsilon > 0$ such that, for all admissible functions x satisfying $\|x - x_*\| \leqslant \varepsilon$, we have $J(x) \geqslant J(x_*)$.

Observe that a strong local minimizer is automatically a weak local minimizer. By extension, we shall say later on that a given property holds in a *strong neighborhood* of x_* provided that, for some $\varepsilon > 0$, it holds at points of the form (t,x) for which $t \in [a,b]$ and $|x - x_*(t)| \leqslant \varepsilon$.

The costate p introduced earlier (p. 310) makes another appearance in expressing the following result.

15.14 Theorem. (Weierstrass c. 1880) *If $x_* \in Lip[a,b]$ is a strong local minimizer for* (P)*, then for almost every $t \in [a,b]$, we have*

$$\Lambda\big(t, x_*(t), x_*'(t) + v\big) - \Lambda\big(t, x_*(t), x_*'(t)\big) \geqslant \langle p(t), v\rangle \;\; \forall\, v \in \mathbb{R}^n.$$

The theorem will be proved in a more general setting later (Theorem 18.1).

When Λ is differentiable, as it is at present, and when $\Lambda(t,x,v)$ is known to be convex with respect to the v variable (which is very often the case in classical problems), the inequality in the theorem is automatically satisfied; the necessary condition of Weierstrass provides no new information. This is because the integral Euler equation implies that the costate p satisfies

$$p(t) = \nabla_v \Lambda(t, x_*(t), x_*'(t)) \text{ a.e.}$$

By convexity therefore, for almost every t, the element $p(t)$ is a subgradient at $x_*'(t)$ of the convex function $v \mapsto \Lambda(t, x_*(t), v)$. The assertion of the theorem is simply the corresponding subgradient inequality (§4.1).

When Λ is nondifferentiable, however, or when it fails to be convex with respect to v (or when additional constraints are present), the Weierstrass necessary condition may furnish significant information about solutions, as we shall see later.

15.15 Example. Consider again the problem discussed in Example 14.14:

$$\min \int_0^1 x'(t)^3\, dt \; : \; x(0) = 0,\; x(1) = 1.$$

We have seen that the arc $x_*(t) = t$ provides a weak local minimum relative to $C^2[0,1]$. Thus, it is a weak local minimizer relative to $\mathrm{Lip}[0,1]$ as well (by Exer. 21.13). However, it is not difficult to see that the Weierstrass condition asserted by Theorem 15.14 cannot hold at any point, since the subdifferential (in the sense of convex analysis) of the function $v \mapsto v^3$ is empty everywhere. We conclude that x_* is *not* a strong local minimizer. □

Chapter 16

Absolutely continuous solutions

The theory of the calculus of variations at the turn of the twentieth century lacked a critical component: it had no existence theorems. These constitute an essential ingredient of the *deductive method* for solving optimization problems, the approach whereby one combines existence, rigorous necessary conditions, and examination of candidates to arrive at a solution. (The reader may recall that in Chapter 9, we discussed at some length the relative merits of the inductive and deductive approaches to optimization.)

The deductive method, when it applies, often leads to the conclusion that a *global* minimum exists. Contrast this, for example, to Jacobi's theorem 14.12, which asserts only the existence of a local minimum. In mechanics, a local minimum is a meaningful goal, since it generally corresponds to a stable configuration of the system. In many modern applications however (such as in engineering or economics), only global minima are of real interest.

Along with the quest for the multiplier rule (which we discuss in the next chapter), it was the longstanding question of existence that dominated the scene in the calculus of variations in the first half of the twentieth century.[1]

16.1 Example. Suppose that we are asked to solve the following instance of the basic problem:

$$\min \int_0^1 \big(1+x(t)\big)\, x'(t)^2\, dt \;:\; x \in C^2[0,1],\, x(0) = 0,\, x(1) = 3.$$

We proceed to apply the necessary conditions for a solution x_*. Since the problem is autonomous, the Erdmann condition (Prop. 14.4) implies

$$\big(1+x_*(t)\big)\, x_*'(t)^2 = c, \quad t \in [0,1]$$

[1] One of Hilbert's famous problems, in the list that he composed in 1900, concerned this issue; another concerned the regularity of solutions.

F. Clarke, *Functional Analysis, Calculus of Variations and Optimal Control*, Graduate Texts in Mathematics 264, DOI 10.1007/978-1-4471-4820-3_16,

for some constant c. If $c=0$, then x_* cannot leave its initial value 0, and thus $x_* \equiv 0$, a contradiction. We deduce therefore that $c>0$. It follows that x_*' is never 0, and therefore $x_*'(t)>0$ for all t, and x_* satisfies

$$\sqrt{1+x_*(t)}\, x_*'(t) \;=\; \sqrt{c}\,, \quad t \in [0,1].$$

We easily solve this separable differential equation and invoke the boundary conditions to find the unique extremal

$$x_*(t) \;=\; (7t+1)^{2/3} - 1,$$

with associated cost $J(x_*) = (14/3)^2$. Having achieved this, we may very well have the impression of having solved the problem.

This is far from being the case, however, since the infimum in the problem is $-\infty$. This can be seen as follows. Consider a function y that is affine between $(0,0)$ and $(1/3,-3)$, and also between $(2/3,-3)$ and $(1,3)$. Between $(1/3,-3)$ and $(2/3,-3)$, we take y to be of "sawtooth" form, with values of y between -4 and -2, and derivative y' satisfying $|y'| = M$ a.e. By taking M increasingly large, we can arrange for $J(y)$ to approach $-\infty$. (The sawtooth function can be approximated in order to achieve the same conclusion using functions in $C^2[a,b]$.)

The serious mistake in the analysis consists of assuming that a solution exists. The purpose of the example is to demonstrate the fallacy of using deductive reasoning when we don't know this to be true.

The sufficient conditions of Theorem 14.12 can be applied to show that x_* is a weak local minimizer; this is an inductive approach. It is also possible (deductively) to show that the function x_* found above *is* the unique global minimizer when the state constraint $x(t) \geqslant 0$ is added to the problem (see Exer. 21.16). □

The key step in developing existence theory is to extend the context of the basic problem to functions that belong to the larger class $\mathrm{AC}[a,b]$ of absolutely continuous functions, rather than $C^2[a,b]$ or $\mathrm{Lip}[a,b]$ as in the preceding sections. Of course, this step could not be taken until Lebesgue had done his great work.

As we did in Chapter 12, we refer to an absolutely continuous function x mapping an interval $[a,b]$ to $\mathbb{R}^n$ as an **arc**; recall that the notation $\mathrm{AC}[a,b]$ is used for arcs on $[a,b]$, even in the vector-valued case. The fact that an arc x has a derivative x' that may be unbounded means that we have to pay some attention to whether $J(x)$ is well defined. (Under our previous hypotheses, this was automatic.)

The phrase "basic problem" now refers to

$$\text{minimize } \; J(x) \,:\, x \in \mathrm{AC}[a,b],\; x(a)=A,\; x(b)=B. \qquad \textbf{(P)}$$

An arc x is **admissible** if it satisfies the constraints of the problem, and if $J(x)$ is well defined and finite. An arc x_* admissible for the basic problem (P) is said to be a solution (or a minimizer) if $J(x_*) \leqslant J(x)$ for all other admissible arcs x.

16.1 Tonelli's theorem and the direct method

The celebrated theorem of Tonelli identifies certain hypotheses under which a solution to (P) exists in the class of arcs. It features a Lagrangian Λ that is continuous and bounded below. Note that, in this case, for any arc x, the function $t \mapsto \Lambda(t, x(t), x'(t))$ is measurable and bounded below. In this setting, then, the integral $J(x)$ is well defined for any arc x, possibly as $+\infty$.

The following result, a turning point in the theory, was the concrete predecessor of the abstract direct method depicted in Theorem 5.51. Note, however, that the functional J is *not* necessarily convex here, which adds to the complexity of the problem.

16.2 Theorem. (Tonelli 1915) *Let the Lagrangian $\Lambda(t,x,v)$ be continuous, convex in v, and* coercive *of degree $r > 1$: for certain constants $\alpha > 0$ and β we have*

$$\Lambda(t,x,v) \geqslant \alpha|v|^r + \beta \;\; \forall\,(t,x,v) \in [a,b] \times \mathbb{R}^n \times \mathbb{R}^n.$$

Then the basic problem (P) *admits a solution in the class $AC[a,b]$.*

Proof. It is clear that there exist admissible arcs x for which $J(x)$ is finite (for example, take x to be the unique affine admissible arc). Accordingly, there exists a minimizing sequence x_i of admissible functions for (P):

$$\lim_{i \to \infty} J(x_i) = \inf(\mathrm{P}) \;\; \text{(finite)}.$$

For all i sufficiently large, in view of the coercivity, we have

$$\int_a^b \big\{ \alpha |x_i'|^r + \beta \big\}\, dt \;\leqslant\; \int_a^b \Lambda\big(t, x_i, x_i'\big)\, dt \;\leqslant\; \inf(\mathrm{P}) + 1.$$

This implies that the sequence x_i' is bounded in $L^r[a,b]^n$. By reflexivity and weak sequential compactness, we may assume without loss of generality that each component converges weakly in $L^r[a,b]$; we label the vector limit v_*.

We proceed to define an element x_* of $\mathrm{AC}^r[a,b]^n$ via

$$x_*(t) = A + \int_a^t v_*(s)\, ds\,, \;\; t \in [a,b].$$

For each $t \in [a,b]$, the weak convergence implies that

$$\int_a^t x_i'(s)\,ds \;=\; \int_a^b x_i'(s)\chi_{(a,t)}(s)\,ds \;\to\; \int_a^b v_*(s)\chi_{(a,t)}(s)\,ds \;=\; \int_a^t v_*(s)\,ds$$

(where $\chi_{(a,t)}$ is the characteristic function of the interval (a,t)), from which we deduce that $x_i(t)$ converges pointwise to $x_*(t)$. (The convergence can be shown to be uniform.) It follows that $x_*(b) = B$, so that x_* is admissible for (P).

We now proceed to invoke the integral semicontinuity theorem 6.38, taking Q to be $[a,b]\times\mathbb{R}^n$, $z_i = x_i'$ and $u_i = x_i$; the convexity of Λ in v is essential here. (Of course, Tonelli does not refer to convex functions in his proof.) We conclude that

$$J(x_*) \;\leqslant\; \lim_{i\to\infty} J(x_i) \;=\; \inf\,(\mathrm{P}).$$

Since x_* is admissible for (P), it follows that x_* is a global minimizer. □

16.3 Exercise. In the following, we outline a more elementary proof of Tonelli's theorem in the case of a "separated" Lagrangian having the form

$$\Lambda(t,x,v) \;=\; f(t,x)+g(t,v).$$

The technicalities are sharply reduced in this case (since no appeal to Theorem 6.38 is required), but the main ideas are the same. We assume (consistently with Theorem 16.2) that f and g are continuous, g is convex in v, f is bounded below, and that we have (for some $\alpha > 0$ and $r > 1$)

$$g(t,v) \;\geqslant\; \alpha|v|^r \;\;\forall\,(t,v) \in [a,b]\times\mathbb{R}^n.$$

(a) Prove that a minimizing sequence x_i exists, and that x_i' is bounded in $L^r(a,b)$.

(b) Prove that the sequence x_i is bounded and equicontinuous, and that, for some subsequence (not relabeled), there is an arc x_* admissible for (P) such that

$$x_i \to x_* \text{ uniformly}, \quad x_i' \to x_*' \text{ weakly in } L^r(a,b).$$

(c) Prove that

$$\int_a^b f\big(t,x_*(t)\big)\,dt \;=\; \lim_{i\to\infty}\int_a^b f\big(t,x_i(t)\big)\,dt.$$

(d) Prove that the mapping

$$v \;\mapsto\; \int_a^b g\big(t,v(t)\big)\,dt$$

is lower semicontinuous on $L^r(a,b)$.

(e) Prove that

$$\int_a^b g\big(t,x_*'(t)\big)\,dt \;\leqslant\; \liminf_{i\to\infty}\int_a^b g\big(t,x_i'(t)\big)\,dt$$

and then conclude that x_* is a solution of (P). □

We proceed now to illustrate how Tonelli's theorem fails if either of the coercivity or the convexity hypotheses is absent. We begin with the convexity.

16.4 Example. Consider the basic problem with $n = 1$,

$$J(x) \;=\; \int_0^1 \big(x(t)^2 + \big[x'(t)^2 - 1\big]^2\big)\,dt\,,$$

and constraints $x(0) = 0,\ x(1) = 0$. Then the Lagrangian $\Lambda(t,x,v)$, which here is the function $x^2 + (v^2-1)^2$, is continuous, and is also coercive of degree 4. We claim that nonetheless, the problem has no solution.

To see this, note that $J(x) > 0$ for any arc x, since we cannot have *both* $x \equiv 0$ and $|\,x'(t)\,| = 1$ a.e. But the infimum in the problem is 0, since, for any positive ε, there is a sawtooth function x whose derivative is ± 1 a.e. and which satisfies $\|x\| < \varepsilon$ (whence $J(x) < \varepsilon^2$).

It is clearly the convexity hypothesis (in v) that is missing here, which is why Tonelli's theorem is inapplicable. Informally, we may say that the sawtooth function "chatters" between the derivative values ± 1 (the locus where $(v^2-1)^2$ attains its minimum) and "almost succeeds" in giving $x = 0$ as well (the locus where x^2 attains its minimum). But complete success is not possible: in the limit, x goes to 0, but x' converges weakly to 0 (see Exer. 6.19). □

The following shows that the degree of coercivity in Tonelli's theorem cannot be lowered to $r = 1$.

16.5 Exercise. Let $g : \mathbb{R} \to \mathbb{R}$ be given by $g(v) \;=\; v\,[\,1+\min\{\,v,0\}\,]$.

(a) Show that g is convex, continuously differentiable, and satisfies

$$g(v) \;\geqslant\; \max\big\{\,v\,,\,|\,v\,|-1\,\big\}\;\;\forall\,v\,.$$

(b) Deduce that any arc x admissible for the problem

$$\min\;\; J(x) = \int_0^1 \big(x^2 + g(x')\big)\,dt \;:\; \text{subject to } x \in \mathrm{AC}[0,1],\; x(0)=0,\; x(1)=1$$

satisfies $J(x) > 1$.

(c) Show that the functions

$$x_i(t) \;=\; \begin{cases} 0 & \text{if } 0 \leqslant t \leqslant 1-1/i \\ i\,[\,t-1+1/i\,] & \text{if } 1-1/i < t \leqslant 1 \end{cases}$$

satisfy $\lim_{i\to\infty} J(x_i) \to 1$.

(d) Conclude that the problem defined in (b) admits no solution. What hypothesis of Tonelli's theorem is missing? □

16.6 Exercise. Let $T > 0$ be fixed. Prove that the following problem has a solution:

$$\text{minimize} \int_0^T \Big\{ \frac{1}{2} m\big(\ell\,\theta'(t)\big)^2 - mg\ell\big(1 - \cos\theta(t)\big)\Big\}\,dt \;: \qquad \theta \in \mathrm{AC}[0,T],\ \theta(0) = 0 = \theta(T).$$

Note that the integral is the action of Example 14.6. Show that if T is large, the solution[2] spends a lot of time near the unstable equilibrium $\theta = \pi$. □

Simple variants of the theorem. The proof of Tonelli's theorem adapts to numerous variants of the basic problem. Some typical ones appear below; they maintain the central elements (coercivity of the Lagrangian, and its convexity in v).

16.7 Exercise. Show that Tonelli's theorem holds when Λ is measurable in t and continuous in (x,v) (rather than continuous in all its variables). An extra hypothesis is now required, however, to ensure the existence of an admissible arc x for which $J(x) < \infty$ (and hence, of a minimizing sequence). A simple one that suffices: for every x and v in $\mathbb{R}^n$, the function $t \mapsto \Lambda(t, x+tv, v)$ is summable. □

16.8 Exercise. Under the same hypotheses as the previous exercise, verify that the proof of Tonelli's theorem is unaffected by the presence in the underlying problem of a *unilateral state constraint*

$$x(t) \in S, \quad t \in [a,b],$$

where S is a given *closed* subset of $\mathbb{R}^n$. Note that in this setting, the coercivity of Λ need only hold when x lies in S. □

16.9 Exercise. Extend Tonelli's theorem to the problem

$$\text{minimize} \quad \ell\big(x(a), x(b)\big) + \int_a^b \Lambda\big(t, x(t), x'(t)\big)\,dt\,, \quad \big(x(a), x(b)\big) \in E\,,$$

where Λ satisfies the hypotheses of the previous exercise, ℓ is continuous, E is closed, and one of the following projections is bounded:

$$\big\{ y \in \mathbb{R}^n : \exists\, z \in \mathbb{R}^n \;:\; (y,z) \in E \big\}, \quad \big\{ z \in \mathbb{R}^n : \exists\, y \in \mathbb{R}^n \;:\; (y,z) \in E \big\}. \quad \square$$

16.10 Exercise. One may extend Tonelli's theorem to certain cases in which the Lagrangian is not necessarily bounded below. A simple instance of this is obtained by weakening the coercivity condition as follows:

$$\Lambda(t,x,v) \;\geqslant\; \alpha|v|^r - \gamma|x|^s + \beta \;\; \forall\,(t,x,v) \in [a,b]\times\mathbb{R}^n\times\mathbb{R}^n,$$

[2] This example shows that the principle of least action does not describe physical reality in the long term, although, as we have seen, it does do so in the short term (and locally).

where $\gamma \geqslant 0$, $0 < s < r$, and where, as before, $r > 1$, $\alpha > 0$. Note that the (positive) growth in v is of higher order than the (possibly negative) growth in x. With this reduced growth condition, the other hypotheses on Λ being unchanged, show that Tonelli's theorem continues to hold. □

The exercise above allows one to assert existence when the Lagrangian is given by $|v|^2 - |x|$, for example.

16.11 Exercise. Let Λ satisfy the hypotheses of Tonelli's theorem, except that the coercivity is weakened to

$$\Lambda(t,x,v) \geqslant \frac{\alpha|v|^r}{(1+|x|)^s} + \beta \;\; \forall\,(t,x,v) \in [a,b] \times \mathbb{R}^n \times \mathbb{R}^n,$$

where $0 < s < r$. Establish the existence of a solution to the basic problem. □

The Direct Method. In view of the variants evoked above, the reader may well feel that some general (albeit ungainly) theorem might be fabricated to cover a host of special cases. Indeed, we could take a stab at this, but experience indicates that circumstances will inevitably arise which will not be covered. It is better to master the method, now known as the *direct method*. (Tonelli, upon introducing it, had called it *a* direct method.)

The underlying approach, then, combines three ingredients: the analysis of a minimizing sequence x_i in order to establish the existence of a subsequence converging in an appropriate sense; the lower semicontinuity of the cost with respect to the convergence; the persistence of the constraints after taking limits. The convergence is generally in the sense that x_i' converges weakly and x_i pointwise or uniformly; the lower semicontinuity typically results from the integral semicontinuity theorem 6.38. To deduce the persistence in the limit of the constraints, the weak closure theorem 6.39 is often helpful. This approach applies to a broad range of problems in dynamic optimization.

In applying the direct method, it is essential to grasp the distinction between constraints that persist under uniform convergence (of x_i) and those that survive weak convergence (of x_i'): convexity is the key to the latter. Consider, for example, the basic problem in the presence of two additional unilateral state and velocity constraints:

$$(a): \;\; x(t) \in S \;\; \forall\, t \in [a,b], \;\text{ and }\; (b): \;\; x'(t) \in V, \;\; t \in [a,b] \text{ a.e.}$$

Uniform (or pointwise) convergence of the sequence x_i will preserve the constraint (a) in the limit, provided only that S is closed. In order for weak convergence to preserve the constraint (b) in the limit, however, we require that V be *convex* as well as closed (as in the weak closure theorem 6.39); compactness does not suffice (see Exer. 8.45). Thus, the appropriate hypotheses for an existence theorem in this context would include: S closed, V closed *and* convex.

As regards the hypotheses giving the weak sequential compactness of the sequence x_i', we would only require the coercivity of Λ to hold for points $x \in S$. The coercivity could also be replaced by compactness of V. (Note that the integral semicontinuity theorem 6.38 does not require a coercive Lagrangian.) Hybrid possibilities can also be envisaged; for example, coercivity of Λ with respect to certain coordinates of v, and compactness of V with respect to the others.

An example of these considerations occurs in the second problem of Exer. 14.23, in which we saw that the problem

$$\text{minimize } \int_0^{\pi} x(t)^2\, dt \quad \text{subject to } \int_0^{\pi} x'(t)^2\, dt = \pi/2$$

does not admit a solution. In order that the isoperimetric constraint

$$\int_a^b \psi\big(t, x(t), x'(t)\big)\, dt = 0$$

be preserved in the limit, ψ needs to be linear with respect to v. On the other hand, the robustness under convergence of a constraint of the form

$$\int_a^b \psi\big(t, x(t), x'(t)\big)\, dt \leqslant 0$$

requires only convexity of ψ with respect to v (as in the integral semicontinuity theorem). In this setting, the required weak compactness could result from various hypotheses. The Lagrangian Λ itself being coercive (as before) would do; but we could require, instead, that the Lagrangian ψ of the isoperimetric constraint be coercive. In general, then, a combination of circumstances will come into play.

16.12 Exercise. Prove that the problem of Exer. 14.23 admits a solution if $C^2[0,\pi]$ is replaced by $\mathrm{AC}[0,\pi]$. (Note, however, that Theorem 14.21 cannot be applied in order to identify it; existence has been achieved, but the necessary conditions have been lost. The analysis is continued in Exer. 17.11.) □

16.2 Regularity via growth conditions

Now that we are armed with an existence theory, we would like to use it in the deductive method, the next step in which is the application of necessary conditions. In examining Theorem 15.2, which asserts the integral Euler equation for the basic problem (P), however, we spot a potential difficulty when absolutely continuous functions are involved. The proof invoked the dominated convergence theorem, which no longer seems to be available when x_*' is unbounded; it appears, in fact, that the differentiability of g cannot be asserted.

Is this problem in proving the existence of $g'(0)$ simply a technical difficulty in adapting the argument, or is it possible that the basic necessary condition for (P) can actually fail? It turns out to be the latter: even for an analytic Lagrangian satisfying the hypotheses of Tonelli's theorem, the integral Euler equation may not hold[3] at the unique minimizing arc x_*. The reader may detect a certain irony here: in order to be able to apply the deductive approach, the basic problem has been extended to $\mathrm{AC}[a,b]$. However, with solutions in this class, the necessary conditions can no longer be asserted.

Another disturbing fact about the extension to arcs is the possibility of the *Lavrentiev phenomenon*. This is said to occur when the infimum in the basic problem over $\mathrm{AC}[a,b]$ is strictly less than the infimum over $\mathrm{Lip}[a,b]$, and it can happen even for smooth Lagrangians satisfying the hypotheses of Tonelli's theorem. From the computational point of view, this is disastrous, since most numerical methods hinge upon minimizing the cost over a class of smooth (hence Lipschitz) functions (for example, polynomials). In the presence of the Lavrentiev phenomenon, such methods cannot approach the minimum over $\mathrm{AC}[a,b]$. The extension from $\mathrm{Lip}[a,b]$ to $\mathrm{AC}[a,b]$, then, is not necessarily a faithful one (a completion), as was the extension from $C^2[a,b]$ to $\mathrm{Lip}[a,b]$.

However, all is not lost. There is a way to recover, in many cases, the happy situation in which we can both invoke existence and assert the necessary conditions, while excluding the Lavrentiev phenomenon. This hinges upon identifying additional structural hypotheses on Λ which serve to rule out the pathological situations cited above. We shall see two important examples of how to do this. The first one below recovers the integral Euler equation under an "exponential growth" hypothesis on the Lagrangian.

Remark. Local minima in the class of arcs are defined essentially as before. For example, an admissible arc x_* is a weak local minimizer if, for some $\varepsilon > 0$, we have $J(x_*) \leqslant J(x)$ for all admissible arcs x satisfying $\|x - x_*\| \leqslant \varepsilon$ and $\|x' - x_*'\| \leqslant \varepsilon$. Recall that the meaning of the word "admissible" in the preceding sentence includes the integral being well defined.

16.13 Theorem. (Tonelli-Morrey) *Let Λ admit gradients Λ_x, Λ_v which, along with Λ, are continuous in (t,x,v). Suppose further that for every bounded set S in $\mathbb{R}^n$, there exist a constant c and a summable function d such that, for all $(t,x,v) \in [a,b] \times S \times \mathbb{R}^n$, we have*

$$\left|\Lambda_x(t,x,v)\right| + \left|\Lambda_v(t,x,v)\right| \leqslant c\big(|v| + \left|\Lambda(t,x,v)\right|\big) + d(t). \qquad (*)$$

Then any weak local minimizer x_* satisfies the Euler equation in integral form.

Proof. It follows from the hypotheses on Λ that the function

[3] The first examples of this phenomenon are quite recent, and exhibit the feature that the function $\Lambda_x(t, x_*(t), x_*'(t))$, which is the derivative of the costate, is not summable.

$$f(t) = \Lambda_v\big(t, x_*(t), x_*'(t)\big) - \int_a^t \Lambda_x\big(s, x_*(s), x_*'(s)\big)\, ds$$

lies in $L^1(a,b)^n$. Let $y : [a,b] \to \mathbb{R}^n$ be a function which belongs to $\operatorname{Lip}[a,b]$, vanishes at a and b, and satisfies $\|y\| + \|y'\| \leqslant 1$. We prove that

$$\int_a^b f(t) \bullet y'(t)\, dt = 0$$

for any such y, which, by Example 9.5, implies the integral Euler equation.

For any $t \in [a,b]$ such that $x_*'(t)$ and $y'(t)$ exist, and such that $|y(t)| + |y'(t)| \leqslant 1$ (thus, for almost every t), we define

$$g(t,s) = \Lambda\big(t, x_*(t) + s y(t), x_*'(t) + s y'(t)\big) - \Lambda\big(t, x_*(t), x_*'(t)\big),\ s \in [0,1].$$

Note that $g(t,0) = 0$ a.e., and, since x_* is a weak local minimizer,

$$\int_a^b g(t,s)\, dt \geqslant 0 \text{ for } s \text{ sufficiently small.} \tag{1}$$

The structural hypothesis $(*)$ yields, for almost every t, for $s \in [0,1]$ a.e.,

$$\begin{aligned} \left| \frac{d}{ds} g(t,s) \right| &\leqslant c\left\{ 1 + |x_*'(t)| + \big|\Lambda(t, x_*(t), x_*'(t))\big| + |g(t,s)| \right\} + d(t) \\ &= c|g(t,s)| + k(t), \end{aligned}$$

for a certain summable function k. This estimate, together with Gronwall's lemma, leads to $|g(t,s)| \leqslant sMk(t)$ for a certain constant M, for all s sufficiently small. In view of this, we may invoke Lebesgue's dominated convergence theorem to deduce (with the help of (1))

$$\begin{aligned} 0 &\leqslant \lim_{s \downarrow 0} \int_a^b \frac{g(t,s) - g(t,0)}{s}\, dt = \int_a^b \frac{d}{ds} g(t,s)\, dt \\ &= \int_a^b \left\{ \Lambda_x\big(t, x_*(t), x_*'(t)\big) \bullet y(t) + \Lambda_v\big(t, x_*(t), x_*'(t)\big) \bullet y'(t) \right\} dt \\ &= \int_a^b f(t) \bullet y'(t)\, dt, \end{aligned}$$

after an integration by parts. Since y may be replaced by $-y$, equality must hold, and the proof is complete. □

16.14 Exercise. Let $\Lambda(t,x,v)$ have the form $f(t,x) + g(v)$, where f and g are continuously differentiable and, for some constant c, the function g satisfies

$$\big|\nabla g(v)\big| \leqslant c\big(1 + |v| + |g(v)|\big) \ \ \forall\, v \in \mathbb{R}^n.$$

Prove that any weak local minimizer for (P) satisfies the integral Euler equation. □

Nagumo growth. The following localized and weakened form of coercivity is useful in regularity theory. It asserts that Λ has superlinear growth in v along x_*.

16.15 Definition. *We say that Λ has Nagumo growth along x_* if there exists a function $\theta : \mathbb{R}_+ \to \mathbb{R}$ satisfying $\lim_{t\to\infty} \theta(t)/t = +\infty$, such that*

$$t \in [a,b], v \in \mathbb{R}^n \implies \Lambda(t, x_*(t), v) \geqslant \theta(|v|).$$

As an illustrative example, we observe that when the Lagrangian satisfies the hypothesis of Exer. 16.10, then Nagumo growth holds along any arc.

16.16 Corollary. *Under the hypotheses of* Theorem 16.13*, if $\Lambda(t,x,v)$ is convex in v and has Nagumo growth along x_*, then x_* is Lipschitz.*

Proof. Under the additional hypotheses, the costate p is the subgradient of Λ in v along x_*, whence

$$\Lambda\big(t, x_*(t), 0\big) - \Lambda\big(t, x_*(t), x_*'(t)\big) \geqslant -p(t) \bullet x_*'(t) \text{ a.e.}$$

Nagumo growth, along with this inequality, reveals:

$$\theta\big(|x_*'(t)|\big) \leqslant \Lambda\big(t, x_*(t), x_*'(t)\big) \leqslant \Lambda\big(t, x_*(t), 0\big) + p(t) \bullet x_*'(t) \text{ a.e.},$$

which implies that x_*' is essentially bounded, since θ has superlinear growth and both $\Lambda(t, x_*(t), 0)$ and $p(t)$ are bounded. □

The desirability of Lipschitz regularity. Note that when the basic problem (P) admits a global solution x_* which is Lipschitz, then the Lavrentiev phenomenon does not occur, and the necessary conditions can be asserted. This is why the *Lipschitz regularity* of the solution is a desirable property. It offers the further advantage of giving us access to the higher regularity results of §15.2, which would allow us to deduce the smoothness of the solution

16.17 Exercise. Use the results above to prove that any solution θ_* to the problem of Exer. 16.6 is Lipschitz. Proceed to show by the results of §15.2 that θ_* is C^∞. □

Remark. Future developments will make it possible to assert the Euler equation and Lipschitz regularity under a weaker growth condition than ($*$) of Theorem 16.13. Specifically, we shall obtain Theorem 16.13 and Cor. 16.16 in §17.3 under the following structural assumption: There exist constants $\varepsilon > 0$ and c, and a summable function d, such that, for almost every t,

$$\big|\Lambda_x\big(t, x, x_*'(t)\big)\big| \leqslant c\big|\Lambda\big(t, x, x_*'(t)\big)\big| + d(t) \;\; \forall x \in B(x_*(t), \varepsilon).$$

16.3 Autonomous Lagrangians

We now prove that under hypotheses of Tonelli type, solutions to the basic problem in the calculus of variations are Lipschitz when the Lagrangian is autonomous. The reader will recall that the problem (P), or its Lagrangian Λ, are said to be *autonomous* when Λ has no dependence on the t variable.

16.18 Theorem. (Clarke-Vinter) *Let $x_* \in AC[a,b]$ be a strong local minimizer for the problem* (P), *where the Lagrangian is continuous, autonomous, convex in v, and has Nagumo growth along x_*. Then x_* is Lipschitz.*

Proof. Let x_* be a solution of (P) relative to $\|x - x_*\| \leqslant \varepsilon$. By uniform continuity, there exists $\delta \in (0,1/2)$ with the property that

$$t, \tau \in [a,b],\ |t-\tau| \leqslant (b-a)\delta/(1-\delta) \implies |x_*(t) - x_*(\tau)| < \varepsilon.$$

A. Let us consider any measurable function $\alpha : [a,b] \to [1-\delta, 1+\delta]$ satisfying the equality $\int_a^b \alpha(t)\,dt = b-a$. For any such α, the relation

$$\tau(t) = a + \int_a^t \alpha(s)\,ds$$

defines a bi-Lipschitz one-to-one mapping from $[a,b]$ to itself; it follows readily that the inverse mapping $t(\tau)$ satisfies

$$\frac{d}{d\tau}\,t(\tau) = \frac{1}{\alpha\big(t(\tau)\big)},\quad |t(\tau)-\tau| \leqslant (b-a)\delta/(1-\delta) \text{ a.e.}$$

Proceed now to define an arc y by $y(\tau) = x_*(t(\tau))$. Then y is admissible for the problem (P), and satisfies $\|y - x_*\| < \varepsilon$ (by choice of δ), whence

$$\int_a^b \Lambda\big(y(\tau), y'(\tau)\big)\,d\tau \geqslant J(x_*).$$

Applying the change of variables $\tau = \tau(t)$ to the integral on the left, and noting that $y'(\tau) = x'(t(\tau))/\alpha(t(\tau))$ a.e., we obtain

$$\begin{aligned}\int_a^b \Lambda\big(y(\tau), y'(\tau)\big)\,d\tau &= \int_a^b \Lambda\big(y(\tau(t)), y'(\tau(t))\big)\,\tau'(t)\,dt \\ &= \int_a^b \Lambda\big(x_*(t), x_*'(t)/\alpha(t)\big)\,\alpha(t)\,dt \geqslant J(x_*).\end{aligned}$$

Note that equality holds when α is the function $\alpha_* \equiv 1$, so we see that α_* solves a certain minimization problem. Let us formulate this problem more explicitly by introducing

$$\Phi(t,\alpha) = \Lambda\big(x_*(t), x_*'(t)/\alpha\big)\,\alpha$$

It is straightforward to verify that for each t, the function $\Phi(t,\cdot)$ is convex on the interval $(0,\infty)$. Consider the functional f given by

$$f(\alpha) = \int_a^b \Phi\big(t,\alpha(t)\big)\,dt\,.$$

Then $f(\alpha)$ is well defined when α is measurable and has values in the interval $[\,1-\delta, 1+\delta\,]$, possibly as $+\infty$, and it follows that f is convex.

For almost every t, by continuity, there exists $\delta(t) \in (0,\delta\,]$ such that

$$\Phi(t,1)-1 \leqslant \Phi(t,\alpha) \leqslant \Phi(t,1)+1 \;\; \forall\,\alpha \in [\,1-\delta(t), 1+\delta(t)\,]\,.$$

It follows from measurable selection theory (§6.2) that we may take $\delta(\cdot)$ measurable.[4] We define S to be the convex subset of $X := L^\infty[\,a,b\,]$ whose elements α satisfy $\alpha(t) \in [\,1-\delta(t), 1+\delta(t)\,]$ a.e.

Consider now an optimization problem (Q) defined on the vector space X. It consists of minimizing f over S subject to the equality constraint

$$h(\alpha) = \int_a^b \alpha(t)\,dt - (b-a) = 0.$$

The argument given above shows that the function $\alpha_* \equiv 1$ solves (Q).

B. We now apply the multiplier rule, more precisely, the version given by Theorem 9.4. We obtain a nonzero vector $\zeta = (\eta,\lambda)$ in $\mathbb{R}^2$ (with $\eta = 0$ or 1) such that

$$\eta f(\alpha) + \lambda h(\alpha) \geqslant \eta f(\alpha_*) \;\; \forall\,\alpha \in S.$$

It follows easily from Theorem 6.31 that $\eta = 1$. Rewriting the conclusion, we have, for any α in S, the inequality

$$\int_a^b \big\{\Lambda\big(x_*(t), x_*'(t)/\alpha(t)\big)\alpha(t) + \lambda\,\alpha(t)\big\}\,dt \geqslant \int_a^b \big\{\Lambda\big(x_*(t), x_*'(t)\big) + \lambda\,\big\}\,dt\,.$$

Invoking Theorem 6.31, we deduce that, for almost every t, the function

$$\alpha \mapsto \theta_t(\alpha) := \Lambda(x_*(t), x_*'(t)/\alpha)\,\alpha + \lambda\,\alpha$$

attains a minimum over the interval $[\,1-\delta(t), 1+\delta(t)\,]$ at the interior point $\alpha = 1$. Fix such a value of t. Then the generalized gradient of θ_t at 1 must contain zero. It follows from nonsmooth calculus (see Exer. 13.23, strangely relevant) that

$$\Lambda\big(x_*(t), x_*'(t)\big) - \langle x_*'(t), \zeta(t)\rangle = -\lambda \quad \text{a.e.,} \tag{1}$$

[4] Because $(0,\delta\,]$ is not closed, there is an implicit (but not difficult) exercise involved here.

where $\zeta(t)$ lies in the subdifferential at $x_*'(t)$ of the convex function $v \mapsto \Lambda(x_*(t), v)$.

C. The last step in the proof is to show (using (1)) that $x_*'(t)$ is essentially bounded. Let t be such that $x_*'(t)$ exists, and such that (1) holds. We have

$$\begin{aligned} &\Lambda\big(x_*(t), x_*'(t)\{1+|x_*'(t)|\}^{-1}\big) - \Lambda(x_*(t), x_*'(t)) \\ &\quad \geqslant \big[\{1+|x_*'(t)|\}^{-1} - 1\big]\langle x_*'(t), \zeta(t)\rangle \text{ (by the subgradient inequality)} \\ &\quad = \big[\{1+|x_*'(t)|\}^{-1} - 1\big]\{\Lambda\big(x_*(t), x_*'(t)\big) + \lambda\}, \end{aligned}$$

in light of (1). Letting M be a bound for all values of Λ at points of the form $(x_*(t), w)$ with $t \in [a,b]$ and $w \in B$, this leads to (in view of the Nagumo growth)

$$\theta\big(|x_*'(t)|\big) \leqslant \Lambda\big(x_*(t), x_*'(t)\big) \leqslant M + \big(M + |\lambda|\big)|x_*'(t)|.$$

The superlinearity of θ implies that $|x_*'(t)|$ is essentially bounded, as required. □

Remark. When Λ is taken to be differentiable in v, then the $\zeta(t)$ that appears in the proof is none other than the costate $p(t) = \Lambda_v(x_*(t), x_*'(t))$, and we see that (1) extends the Erdmann condition (see Prop. 14.4) to the current setting (with $h = \lambda$). It has now been obtained for x_* merely Lipschitz rather than C^2; the simple proof used before no longer pertains.

The reader may verify that the coercivity was not used to obtain the Erdmann condition, and that its proof goes through unchanged in the presence of an explicit state constraint $x(t) \in S$ in the problem (P), and also when a constraint $x'(t) \in C$ is imposed, provided that C is a cone. We summarize these observations:

16.19 Corollary. *Let Λ be continuous and autonomous and, with respect to v, be convex and differentiable. Let x_* be a strong local minimizer for the problem*

$$\text{minimize } J(x) : x \in AC[a,b],\ x(t) \in S,\ x'(t) \in C,\ x(a) = A,\ x(b) = B,$$

where S is a subset of $\mathbb{R}^n$ and C is a cone in $\mathbb{R}^n$. Then, for some constant h, the arc x_* satisfies the **Erdmann condition**

$$\langle x_*'(t), \Lambda_v(x_*(t), x_*'(t))\rangle - \Lambda(x_*(t), x_*'(t)) = h \text{ a.e.}$$

If in addition Λ has Nagumo growth along x_*, then x_* is Lipschitz.

The reader will notice that as a result of the above, and in contrast to Chapter 14, the Erdmann condition is now available as a separate necessary condition for optimality in certain situations in which the Euler equation cannot be asserted (because of the additional constraints, or simply because x_* is not known to be Lipschitz). Exers. 21.15 and 21.16 illustrate its use in such situations.

16.20 Exercise. Consider the problem of Exer. 16.6. Letting its solution be θ_*, show that the Erdmann condition asserts

$$\frac{1}{2}m\big(\ell\,\theta_*'(t)\big)^2 + mg\ell\big(1-\cos\theta_*(t)\big) \;=\; h.$$

Note that this corresponds to *conservation of energy*, which is often the interpretation of the Erdmann condition in classical mechanics. □

16.21 Example. We illustrate now the use of the existence and regularity theorems, and also their role in studying boundary-value problems in ordinary differential equations. Consider the following version of the basic problem (P), with $n = 1$:

$$\min \int_0^T \left\{ \frac{x(t)^4}{4} - \frac{x(t)^2}{2} + \frac{x'(t)^2}{2} \right\} dt \;:\; x \in \mathrm{AC}[0,T],\; x(0) = 0,\; x(T) = 0.$$

The Lagrangian

$$\Lambda(x,v) \;=\; v^2/2 + x^4/4 - x^2/2$$

is continuous and convex in v, and (it can easily be shown) coercive of degree 2. According to Tonelli's theorem, there exists a solution x_* of (P).

It follows now from Theorem 16.18 that x_* is Lipschitz, since Λ is autonomous. An alternative to calling upon Theorem 16.18 is to argue as follows. We have

$$\frac{|\Lambda_x| + |\Lambda_v|}{1+|v|+|\Lambda(x,v)|} \;\leqslant\; \frac{|v|+|x|^3+|x|}{1+|v|} \;\leqslant\; 1+|x|^3+|x|,$$

which shows that the structural hypothesis ($*$) of Theorem 16.13 holds. This allows us to invoke Cor. 16.16 in order to conclude that x_* is Lipschitz.

In either case, it follows that x_* satisfies the integral Euler equation. We then appeal to Theorem 15.7 in order to deduce that $x_* \in C^\infty[0,T]$. This allows us to write the Euler equation in fully differentiated form:

$$x''(t) \;=\; x(t)\big(x(t)^2 - 1\big).$$

In summary, there is a solution x_* of the boundary-value problem

$$x''(t) \;=\; x(t)\big(x(t)^2 - 1\big),\;\; x \in C^\infty[0,T],\; x(0) = 0,\; x(T) = 0, \tag{2}$$

one that also solves the problem (P).

However, it is clear that the zero function is a solution of (2), and we would wish to know when there is a *nontrivial* solution. This will certainly be the case if the zero function, which is evidently an extremal for Λ, admits a conjugate point in the interval $(0,T)$. For in that case, it cannot be a solution of (P), by the necessary condition of Jacobi (Theorem 14.12), whence $x_* \not\equiv 0$.

The Jacobi equation (for the zero function) is $u''(t) + u(t) \;=\; 0$, which yields the conjugate point $\tau \;=\; \pi$. We arrive therefore at the following conclusion: there exists a nontrivial solution of (2) when $T \;>\; \pi$. □

16.22 Exercise. We consider the following problem (P):

$$\text{minimize} \int_0^1 \exp\{\, x(t) + x'(t)^2 \,\}\, dt \;:\; x \in \mathrm{AC}[0,1],\ x(0) = 0,\ x(1) = 1\,.$$

(a) Use the direct method to prove that (P) admits a solution.

(b) Prove that (P) admits a unique solution.

(c) Observe that the Lagrangian does *not* satisfy hypothesis ($*$) of Theorem 16.13.

(d) Prove that the solution of (P) is Lipschitz.

(e) Deduce the existence of a *unique* solution x to the following boundary-value problem:

$$x''(t) = \frac{1 - x'(t)^2}{1 + x'(t)^2}\,,\quad x \in C^\infty[0,1],\ x(0) = 0,\ x(1) = 1\,.$$

□

Remark. We are now able to reflect with hindsight on the role of each of the three different function spaces that have figured in the theory. The choice of $C^2[a,b]$ is agreeable for evident reasons of simplicity and smoothness. We venture on to $\mathrm{Lip}[a,b]$ because this space still leads to a good theory, including the basic necessary conditions, while allowing nonsmooth solutions; further, there are regularity results that establish a bridge back to $C^2[a,b]$. Finally, we advance to $\mathrm{AC}[a,b]$ because it makes existence theorems possible; and again, there exist bridges from $\mathrm{AC}[a,b]$ that lead back to $\mathrm{Lip}[a,b]$ in many cases. [5]

[5] Note that the class PWS of piecewise-smooth functions would not fit into this scheme: there are no bridges from $\mathrm{AC}[a,b]$ to PWS.

Chapter 17
The multiplier rule

The reader has been told that the great twentieth-century quests in the calculus of variations have involved existence and multiplier rules. Progress in functional analysis, together with the direct method, has largely resolved the former issue; we turn now to that of multipliers.

For this purpose, we consider the classical *problem of Lagrange*. It consists of the basic problem (P) to which has been grafted an additional pointwise equality constraint $\varphi(t,x,x') = 0$, where φ has values in $\mathbb{R}^k$; thus the problem becomes

$$\left\{\begin{array}{ll} \text{Minimize} & J(x) = \displaystyle\int_a^b \Lambda\big(t, x(t), x'(t)\big)\,dt \\ \text{subject to} & \\ & x \in \mathrm{AC}[a,b],\ x(a) = A,\ x(b) = B \\ & \varphi\big(t, x(t), x'(t)\big) = 0,\ t \in [a,b] \text{ a.e.} \end{array}\right. \qquad (\mathbf{L})$$

As before, the arcs x have values in $\mathbb{R}^n$. The additional constraint makes the problem (L) much more complex than (P), or even the isoperimetric problem. In part, this is because we now have infinitely many constraints, one for each t.

Given our experience in optimization, it is not hard to fathom the general nature of the necessary conditions we seek. We expect the multiplier rule to assert, in a now familiar pattern, that if x_* solves this problem, then there exist multipliers η, λ, not both zero, with $\eta = 0$ or 1, such that x_* satisfies (some or all of) the necessary conditions for the Lagrangian $\eta\Lambda + \lambda(t) \bullet \varphi$. Note that λ is a function of t here, which is to be expected, since there is a constraint $\varphi(t,x,x') = 0$ for each t.

We are not surprised that a result of this type requires non degeneracy of the constraint. More explicitly, a *rank condition* is postulated: it requires that $D_v\,\varphi(t,x,v)$ have rank k at relevant points (thus, $k \leqslant n$ necessarily).

There are various results of this type, and proving any of them is an arduous task, even if one is generous with hypotheses. But prove them we will, later on, in a more general context. In this chapter, we proceed to describe some representative theorems of multiplier rule type, in an attempt to convey to the reader the central issues. Inevitably, one of these is regularity: in which class is the solution?

F. Clarke, *Functional Analysis, Calculus of Variations and Optimal Control*, Graduate Texts in Mathematics 264, DOI 10.1007/978-1-4471-4820-3_17,

17.1 A classic multiplier rule

Let x_* be a weak local minimizer for the problem (L) above, where we take φ and Λ to be continuously differentiable. If x_* is also smooth, as it is in the classical setting, then the rank condition can be postulated in a pointwise manner. Accordingly, we make the assumption that the solution x_* is piecewise C^1. Let x_* have one of its (finitely many) corners at $\tau \in (a,b)$; we denote by $x_*'(\tau+)$ and $x_*'(\tau-)$ its derivatives from the right and from the left at τ (these exist by assumption).

The **rank condition** in this context is the following:

The Jacobian matrix $D_v\,\varphi(t,x_*(t),x_*'(t))$ has rank k for every $t \in [a,b]$, where, if x_* has a corner at τ, this holds with both the one-sided derivatives $x_*'(\tau+)$ and $x_*'(\tau-)$ in the place of $x_*'(\tau)$.

17.1 Theorem. *Under the hypotheses above, there exist η equal to 0 or 1, an arc p, and a bounded measurable function $\lambda : [a,b] \to \mathbb{R}^k$ such that*

$$\big(\eta, p(t)\big) \neq 0 \;\; \forall\, t \in [a,b],$$

and such that x_* satisfies the **Euler equation:**

$$\big(p'(t), p(t)\big) \;=\; \nabla_{x,v}\left\{\eta\Lambda + \langle \lambda(t), \varphi\rangle\right\}\big(t, x_*(t), x_*'(t)\big) \text{ a.e.}$$

Observe that the theorem asserts the integral form of the Euler equation, as expressed through the costate p, for the Lagrangian $\eta\Lambda + \langle \lambda(t), \varphi\rangle$. The proof (given in Cor. 25.16) will show that we can also add a *local* Weierstrass condition to this classical result: for some $\delta > 0$, for every non corner point t of x_*, we have

$$|v - x_*'(t)| < \delta,\; \varphi(t,x_*(t),v) = 0 \Longrightarrow \\ \eta\,\Lambda\big(t,x_*(t),v\big) - \eta\,\Lambda\big(t,x_*(t),x_*'(t)\big) \geqslant \langle p(t), v - x_*'(t)\rangle\,.$$

17.2 Example. Let us revisit the problem of the hanging chain (see Example 14.22), in the case in which the chain is *not* homogeneous. Let the mass density be $\sigma(y)$ at the point y along the length of the chain, $0 \leqslant y \leqslant L$. Letting $x(\cdot)$ be the shape of the chain at equilibrium (as before), the potential energy is given by

$$\int_a^b \sigma\Big(\int_a^t \sqrt{1+x'(t)^2}\,ds\Big)\,x(t)\,\sqrt{1+x'(t)^2}\,dt\,.$$

We seek to minimize this functional subject to

$$x(a) = A,\;\; x(b) = B,\;\; \int_a^b \sqrt{1+x'(t)^2}\,dt \;=\; L\,.$$

Note that when the function σ is nonconstant, this does *not* have the form of the basic problem, and, on the face of it, we do not know how to proceed. However, let us introduce a new state variable $y(t)$ to stand in for the argument of σ in the integral above; that is, y must satisfy

$$y'(t) = \sqrt{1+x'(t)^2}\,,\; y(a) = 0,\; y(b) = L.$$

Then the problem may be rephrased as follows:

$$\begin{cases} \text{minimize } \displaystyle\int_a^b \sigma\big(y(t)\big)\,x(t)\,\sqrt{1+x'(t)^2}\;dt \quad \text{subject to} \\ \big(x(a),y(a)\big) = (A,0),\; \big(x(b),y(b)\big) = (B,L),\; \varphi\big(x(t),y(t),x'(t),y'(t)\big) = 0, \end{cases}$$

where $\varphi(x,y,v,w) = w - \sqrt{1+v^2}$. This has the form of the basic problem, with state (x,y), and in the presence of an additional pointwise constraint specified by φ; a problem of Lagrange, *quoi*.

Observe that the rank condition is satisfied, since

$$D_{v,w}\,\varphi(x,y,v,w) = \big(-v/\sqrt{1+v^2}\,,\,1\big)$$

always has rank one. If the multiplier rule above is invoked, it leads to the augmented Lagrangian

$$\eta\,\sigma(y)\,x\sqrt{1+v^2} + \lambda(t)\big(w - \sqrt{1+v^2}\,\big).$$

If $\eta = 0$, the extremals correspond to affine functions. This abnormal case may be ruled out; thus, we take $\eta = 1$. We proceed to write the integral Euler equation: there exist arcs p and q such that

$$p' = \sigma(y)\sqrt{1+x'^2} \qquad p = \frac{\big(\sigma(y)\,x-\lambda\big)\,x'}{\sqrt{1+x'^2}}$$

$$q' = \sigma'(y)\,x\sqrt{1+x'^2} \qquad q = \lambda\,.$$

Introducing an additional state v satisfying $x' = v$, we may rewrite this (after some fiddling) as a first-order system; we obtain the following boundary-value problem for four functions (x,v,y,λ):

$$x' = v \qquad v' = \sigma(y)\big(1+v^2\big)/\big(\sigma(y)\,x-\lambda\big)$$

$$y' = \sqrt{1+v^2} \qquad \lambda' = \sigma'(y)\,x\sqrt{1+v^2}$$

$$\big(x(a),y(a)\big) = (A,0),\; \big(x(b),y(b)\big) = (B,L).$$

We are now in the (easily-reached) zone where explicit closed-form solutions are unlikely; the boundary-value problem itself must be viewed as the "answer." It is easy to verify, however, that this reduces to the case considered in Example 14.22 when σ is constant (the homogeneous chain), in which case x is a catenary. □

The following exercise reveals another context in which problems of Lagrange arise: functionals in which occur higher derivatives of the state. (Such problems are common in elasticity.)

17.3 Exercise. Let x_* minimize the functional

$$\int_0^1 \Lambda\big(t, x(t), x'(t), x''(t)\big)\, dt$$

subject to the constraints

$$x \in \mathrm{AC}[0,1],\ x' \in \mathrm{AC}[0,1],\ x(0) = x_0,\ x(1) = x_1,\ x'(0) = v_0,\ x'(1) = v_1,$$

where the Lagrangian $\Lambda(t, x, v, w)$ lies in C^2. We suppose that $x_* \in C^3[0,1]$. To treat this situation, let us introduce an additional state variable y together with the constraint $y - x' = 0$. Show that when rephrased in this way, the problem becomes one to which Theorem 17.1 can be applied. Go on to show that, in doing so, $\eta = 1$ necessarily. Then deduce that x_* satisfies the *second-order Euler equation*

$$-\frac{d^2}{dt^2}\big\{\Lambda_w(*)\big\} + \frac{d}{dt}\big\{\Lambda_v(*)\big\} = \Lambda_x(*),\ t \in [0,1],$$

where $(*)$ refers to evaluation at the point $(t, x_*(t), x_*'(t), x_*''(t))$. □

17.2 A modern multiplier rule

A more general multiplier rule than that of Theorem 17.1 would be one that reduces the assumed regularity of the solution, asserts other necessary conditions in addition to the Euler equation, and offers the possibility of treating other types of pointwise constraints, notably inequalities like $\varphi(t, x, x') \leqslant 0$. We record now a multiplier rule that incorporates these extensions, for the following problem:

$$\left\{\begin{array}{ll} \text{Minimize} & J(x) = \displaystyle\int_a^b \Lambda\big(t, x(t), x'(t)\big)\, dt \\ \text{subject to} & \\ & x \in \mathrm{AC}[a,b],\ \big(x(a), x(b)\big) \in E \\ & \varphi\big(t, x(t), x'(t)\big) \in \Phi,\ t \in [a,b] \text{ a.e.} \end{array}\right. \qquad (\mathbf{L}')$$

where $E \subset \mathbb{R}^n \times \mathbb{R}^n$ and $\Phi \subset \mathbb{R}^k$ are closed sets. We take the functions φ and Λ to be continuously differentiable.

Note that (L′) extends the problem (L) (p. 335) by allowing pointwise constraints of a more general nature, and also more general boundary conditions.

Let x_* be a Lipschitz arc providing a strong local minimum for problem (L$'$). We shall posit the following **rank condition** (or constraint qualification), which is taken to hold in a strong neighborhood of x_*: for some $\varepsilon > 0$,

$$t \in [a,b],\ |x - x_*(t)| \leqslant \varepsilon,\ \varphi(t,x,v) \in \Phi,\ \lambda \in N^L_\Phi(\varphi(t,x,v)),\\ 0 = D_v\langle \lambda, \varphi\rangle(t,x,v) \implies \lambda = 0.$$

The reader may verify that when $\Phi = \{0\}$ (as in the problem (L) considered earlier), this is equivalent to requiring that $D_v\varphi$ have rank k at (nearby) points for which $\varphi = 0$, a familiar type of hypothesis. For the case of inequality constraints (that is, when $\Phi = \mathbb{R}^k_-$), we would obtain a condition along the lines of positive linear independence.

17.4 Theorem. *Under the hypotheses above, there exist η equal to 0 or 1, an arc p, and a bounded measurable function $\lambda : [a,b] \to \mathbb{R}^k$ with*

$$\big(\eta, p(t)\big) \neq 0\ \ \forall t \in [a,b],\ \ \lambda(t) \in N^L_\Phi\big(\varphi(t, x_*(t), x_*'(t))\big) \text{ a.e.}$$

such that x_* satisfies the **Euler equation:**

$$\big(p'(t), p(t)\big) = \nabla_{x,v}\big\{\eta\Lambda + \langle \lambda(t), \varphi\rangle\big\}\big(t, x_*(t), x_*'(t)\big) \text{ a.e.},$$

the **Weierstrass condition:** *for almost every t,*

$$v \in \mathbb{R}^n,\ \varphi\big(t, x_*(t), v\big) \in \Phi \implies\\ \eta\Lambda\big(t, x_*(t), v\big) - \eta\Lambda\big(t, x_*(t), x_*'(t)\big) \geqslant \langle p(t), v - x_*'(t)\rangle,$$

as well as the **transversality condition:**

$$\big(p(a), -p(b)\big) \in N^L_E\big(x_*(a), x_*(b)\big).$$

Notice that the theorem assumes *a priori* that x_* is Lipschitz. In some cases, this would be guaranteed by the Lagrange constraint itself, if it so happens that the condition $\varphi(t, x(t), x'(t)) \in \Phi$ forces x' to be bounded.

The proof of a multiplier rule such as this is far from simple, and, in fact, we are not prepared to give one now. Theorem 17.4 will follow from later results (see Cor. 25.15). Let us proceed nonetheless to illustrate its use.

17.5 Example. We consider the problem

$$\text{minimize} \int_0^3 \big\{x'(t)^2 + 4x(t)\big\}\,dt : x'(t) \geqslant -2 \text{ a.e.},\ x(0) = 0 = x(3),$$

where $x \in \mathrm{AC}[0,3]$ and $n = 1$. Let us assume that a Lipschitz solution x_* exists. The problem can be put into the form of the problem (L$'$) of Theorem 17.4 by taking

$$\Lambda(t,x,v) = v^2+4x, \quad \varphi(t,x,v) = -v-2, \quad \Phi = (-\infty,0].$$

The rank condition of Theorem 17.4 is easily seen to hold, since we have here $D_v\langle\lambda,\varphi\rangle = -\lambda$. We may now invoke the theorem to deduce the existence of η, λ, and p as described. The condition $\lambda(t) \in N^L_\Phi(\varphi(t,x_*(t),x_*'(t)))$ translates in the current setting as follows: for almost every t,

$$\lambda(t) \geqslant 0, \text{ and } \lambda(t) = 0 \text{ if } x_*'(t) > -2.$$

We may exclude the abnormal case $\eta = 0$. For then we would have $p(t) \neq 0\ \forall\, t$, whence $\lambda(t) \neq 0$ a.e. (from the Euler equation), so that the inequality constraint is saturated: $x_*' = -2$ a.e. But this is inconsistent with $x_*(0) = x_*(3) = 0$. Setting $\eta = 1$, therefore, the Euler equation becomes

$$p' = 4, \quad p = 2x_*' - \lambda \text{ a.e.}$$

Thus p is strictly increasing. We now examine the Weierstrass condition. This asserts that, for almost every t, the value $x_*'(t)$ minimizes the function $v \mapsto v^2 - p(t)v$ subject to $v \geqslant -2$. This implies

$$x_*'(t) = \begin{cases} -2 & \text{if } p(t) \leqslant -4 \\ p(t)/2 > -2 & \text{if } p(t) > -4 \end{cases}$$

which, in turn, implies that x_*' is continuous on $(0,3)$.

It follows from this analysis that there is a point $\tau \in [0,3)$ having the property that

$$t < \tau \implies x_*'(t) = -2, \quad t > \tau \implies x_*'(t) > -2.$$

On the interval $[\tau,3]$, we have $\lambda = 0$ a.e. and $x_*'' = 2$. If $\tau = 0$, then x_* is the function $t^2 - 3t$, which cannot be the case, since this function violates the constraint $x'' \geqslant -2$. Thus $\tau > 0$.

To summarize to this point, we know that x_* is continuously differentiable on $(0,3)$ and takes the form

$$x_*(t) = \begin{cases} -2t & \text{if } t \leqslant \tau \\ (t-3)^2 + c(t-3) & \text{if } t \geqslant \tau \end{cases}$$

for some $\tau \in (0,3)$ and constant c. The continuity of both x_* and x_*' at τ provides two equations for τ and c; we find $\tau = 3-\sqrt{6},\ c = 2(\sqrt{6}-1)$.

The conclusion of our analysis is that if a Lipschitz solution exists, it can only be the arc x_* that we have identified. We continue the analysis later by showing that, in fact, x_* *does* provide a global minimum for the problem (see Example 18.9). □

The classical state constraint. As we saw in connection with geodesics, certain problems are such that the competing arcs are naturally restricted to a surface S defined by $\psi(x) = 0$. We discuss such a case now, for the problem

$$\begin{cases} \text{Minimize} & J(x) = \int_a^b \Lambda\big(t, x(t), x'(t)\big)\, dt \\ \text{subject to} & \\ & x \in \mathrm{AC}[a,b],\ x(a) = A,\ x(b) \in E_1 \\ & x(t) \in S = \{u \in \mathbb{R}^n : \psi(x) = 0\},\ t \in [a,b]. \end{cases} \tag{$\mathbf{L}''$}$$

Here, $\psi : \mathbb{R}^n \to \mathbb{R}^k$ is taken to be C^2, with $1 \leqslant k < n$, and Λ is continuously differentiable; E_1 is closed.

The problem only makes sense if $\psi(A) = 0$, since admissible arcs x begin at A; this leads to an evident reduction: an arc x will satisfy the initial condition and remain in S if and only if it satisfies

$$\frac{d}{dt}\,\psi\big(x(t)\big) = D\psi\big(x(t)\big)x'(t) = 0 \text{ a.e.}$$

Thus, the state constraint may be replaced by the Lagrange condition $\varphi(x, x') = 0$, where $\varphi(x, v)$ is defined to be $D\psi(x)v$. Then we may seek to apply Theorem 17.4, whose rank hypothesis concerns $D_v\varphi(x, v) = D\psi(x)$. Now it is quite reasonable to assume that this has maximal rank k at points in S (as we do). Indeed, this **classical rank condition** is a customary one when considering manifolds, and we posit it for our purposes. In its presence, one can justify invoking Theorem 17.4 in order to express necessary conditions.

There is a complication, however, as we proceed to explain. Let x be *any* arc admissible for (L$''$), and define

$$p(t) = \nabla\langle \lambda, \psi\rangle\big(x(t)\big),$$

where λ is any nonzero constant vector in $\mathbb{R}^k$. The reader may verify that, with η taken to be 0, and because of the particular structure of φ, this costate p satisfies the nontriviality condition, the Euler equation, and the Weierstrass condition of Theorem 17.4. If the remaining necessary condition, the transversality condition, always holds as well (which is certainly the case when E_1 is a singleton), then we must acknowledge that the necessary conditions obtained in this way are *trivial* (since they are satisfied by any admissible arc).

This explains the requirement for some extra hypothesis that serves to exclude the triviality described above. For this purpose, we introduce the following *endpoint compatibility condition*:

$$N_S^L(x) \cap N_{E_1}^L(x) = \{0\}\ \ \forall x \in E_1 \cap S.$$

Besides excluding triviality, it will allow us to state necessary conditions in normal form. Note that $N_S^L(x) = N_S(x)$ here, since S is a classical manifold.

When $E_1 = \mathbb{R}^n$, so that the right endpoint is free, then the endpoint compatibility condition certainly holds, since the normal cone $N_{E_1}^L$ reduces to $\{0\}$. At the other extreme, when E_1 is a singleton $\{x_1\}$, the compatibility condition clearly fails, since $N_{E_1}^L(x_1)$ is the whole space and $N_S^L(x_1)$ is a subspace of dimension $n-k \geqslant 1$ (by Cor. 5.37). However, we shall see below in a corollary how to recover this important special case.

17.6 Theorem. *Let $x_* \in Lip[a,b]$ be a strong local minimizer for problem* (L″), *under the rank and compatibility hypotheses above. Set $\phi(x,v) = D\psi(x)v$. Then there exist an arc p and a bounded measurable function $\lambda : [a,b] \to \mathbb{R}^k$ such that x_* satisfies the* **Euler equation:**

$$\big(p'(t), p(t)\big) = \nabla_{x,v}\big\{\Lambda + \langle \lambda(t), \varphi\rangle\big\}\big(t, x_*(t), x_*'(t)\big) \text{ a.e.},$$

the **Weierstrass condition:** *for almost every t,*

$$v \in \mathbb{R}^n,\ D\psi(x_*(t))v = 0 \implies$$
$$\Lambda\big(t, x_*(t), v\big) - \Lambda\big(t, x_*(t), x_*'(t)\big) \geqslant \langle p(t), v - x_*'(t)\rangle,$$

as well as the **transversality condition:**

$$-p(b) \in N_{E_1}^L\big(x_*(b)\big).$$

Proof. As pointed out above, the state constraint $\psi(x(t)) = 0$ is equivalent to

$$\frac{d}{dt}\,\psi(x(t)) = \phi(x(t), x'(t)) = 0 \text{ a.e.},$$

where $\varphi(x,v) = D\psi(x)v$. It is a routine exercise to show that the matrix $D\psi(x)$ has maximal rank in a strong neighborhood of x_*; this implies the rank hypothesis of Theorem 17.4 bearing upon $D_v\varphi(x,v) = D\psi(x)$. Thus, we may apply Theorem 17.4. It remains to show, however, that $\eta \neq 0$. We do so by contradiction. If $\eta = 0$, then

$$p(t) = \nabla\langle \lambda(t), \psi\rangle\big(x_*(t)\big) \text{ a.e.},$$

and transversality yields

$$-p(b) = -\nabla\langle \lambda, \psi\rangle\big(x_*(b)\big) \in N_{E_1}^L\big(x_*(b)\big),$$

for some essential cluster point λ of $\lambda(\cdot)$ at b. But the point $-\nabla\langle\lambda,\psi\rangle(x_*(b))$ is an element of $N_S(x_*(b))$ (see Theorem 5.35), so by the compatibility condition it must equal zero. It follows from this, together with the rank condition, that $\lambda = 0$. Then $p(b) = 0$, and the nontriviality condition of Theorem 17.4 is violated. This is the required contradiction. □

17.7 Corollary. *The theorem continues to hold when E_1 is a singleton $\{x_1\}$.*

Proof. Define $\widetilde{E}_1 = (x_1 + N_S(x_1)) \cap B(x_1, r)$, where $r > 0$ has the property that the intersection of the set so defined with S is the singleton $\{x_1\}$; such r exists by Prop. 1.43. Then we obtain the same problem (locally) if the endpoint constraint is replaced by $x(b) \in \widetilde{E}_1$. But this modified problem satisfies the compatibility condition, since

$$N_{\widetilde{E}_1}(x_1) = N_S(x_1)^{\triangle} = T_S(x_1)$$

(because $N_S(x_1)$ is a subspace) and since $T_S(x_1) \cap N_S(x_1) = \{0\}$. Then we may apply Theorem 17.6, which gives the required conclusion. □

17.8 Exercise. A particle of mass m is restricted to the x-y plane, more precisely to the curve $y = x^2$, and is acted upon only by gravity. The principle of least action asserts that (for small time intervals) the path $(x(t), y(t))$ of the particle will minimize the action; accordingly, we consider, for $T > 0$ fixed, the problem

$$\min \int_0^T \left\{ \frac{1}{2} m \big| \big(x'(t), y'(t)\big) \big|^2 - mgy(t) \right\} dt$$

subject to the state constraint

$$\psi\big(x(t), y(t)\big) = x(t)^2 - y(t) = 0, \ t \in [0, T].$$

The values of the state are prescribed at 0 and T.

(a) Prove that a solution (x, y) of the minimization problem exists.

(b) Invoke Cor. 16.19 to deduce that the solution is Lipschitz.

(c) Verify that Cor. 17.7 applies. Use the Weierstrass condition to prove that x' is absolutely continuous; go on to deduce that y' is absolutely continuous.

(d) Deduce from the Euler equation that the multiplier $\lambda(t)$ is absolutely continuous.

(e) Obtain the governing dynamics for $x(t)$:

$$x''(t) = \frac{-2x(t)[g + 2x'(t)^2]}{1 + 4x(t)^2}.$$

□

Remark. There is a theory of necessary conditions for problems incorporating inequality state constraints of the form $g(x(t)) \leqslant 0$. The distinction with an equality constraint is that the inequality may only be saturated on certain subintervals. It turns out that the costate $p(t)$ is typically discontinuous at points t for which the state constraint becomes, or ceases to be, active. For that reason, the results for such problems, which we do not consider here, are usually phrased in terms of costates p of *bounded variation*.

17.3 The isoperimetric problem

In both the classical (Theorem 17.1) and modern (Theorem 17.4) forms of the multiplier rule that we have seen, an *a priori* assumption regarding the solution x_* was made (piecewise smooth or Lipschitz).

It is possible to do away with such assumptions if additional structural hypotheses are made concerning the Lagrangian. We illustrate this now in the case of the problem

$$\text{minimize} \quad J(x) = \int_a^b \Lambda\big(t, x(t), x'(t)\big)\, dt$$

subject to boundary conditions $x(a) = A$, $x(b) = B$, and the isoperimetric constraint

$$\int_a^b \psi(t, x(t), x'(t)) = 0,$$

where ψ has values in $\mathbb{R}^k$. The functions ψ and Λ are continuously differentiable. For a given strong local minimizer $x_* \in \mathrm{AC}[a,b]$, here is the *structural growth hypothesis* that we postulate:

There exist $\varepsilon > 0$, a constant c, and a summable function d such that, for almost every t,

$$\big|D_x\big(\Lambda, \psi\big)\big(t, x, x_*'(t)\big)\big| \leqslant c\,\big|\big(\Lambda, \psi\big)(t, x, x_*'(t))\big| + d(t) \;\; \forall x \in B(x_*(t), \varepsilon).$$

The reader will observe that this condition is automatically satisfied when x_* is Lipschitz, since in that case the left side admits a uniform bound.

The norm of a Jacobian matrix appeared in this last inequality; perhaps this is a good time to mention:

Notation. The norm $|M|$ of an $m \times n$ matrix M is defined to be the Euclidean norm of its entries viewed as an element of $\mathbb{R}^{mn}$.

We remark that with this choice, and viewing points in $\mathbb{R}^n$ as columns, we obtain the general inequality $|Mu| \leqslant |M||u|$.

17.9 Theorem. *Under the hypotheses above, there exist η equal to 0 or 1, an arc p, and $\lambda \in \mathbb{R}^k$ with $(\eta, \lambda) \neq 0$ such that x_* satisfies the* **Euler equation:**

$$\big(p'(t), p(t)\big) = \nabla_{x,v}\big\{\eta\Lambda + \langle \lambda, \psi \rangle\big\}\big(t, x_*(t), x_*'(t)\big) \text{ a.e.}$$

and the **Weierstrass condition***: for almost every t, for every $v \in \mathbb{R}^n$,*

$$\big(\eta\Lambda + \langle \lambda, \psi \rangle\big)\big(t, x_*(t), v\big) - \big(\eta\Lambda + \langle \lambda, \psi \rangle\big)\big(t, x_*(t), x_*'(t)\big) \geqslant \langle p(t), v - x_*'(t) \rangle.$$

Remark. Note that no rank condition is imposed here. Thus, nothing prevents us from taking ψ to be identically zero. In that case, the growth condition of the theorem bears upon Λ alone, and constitutes a considerable weakening of the Tonelli-Morrey structural hypothesis $(*)$ of Theorem 16.13.

Furthermore, in contrast to the classical isoperimetric multiplier rule, the Weierstrass condition now accompanies the Euler equation in the conclusion. We emphasize that the theorem (which is proved later as a consequence of more general results; see Cor. 22.18) neither requires that x_* be Lipschitz nor asserts that it is.

17.10 Example. The following isoperimetric problem arises in the study of periodic Hamiltonian trajectories: to minimize

$$\int_0^1 \Lambda\left(-y', x'\right) dt \quad \text{subject to} \quad \int_0^1 y \bullet x' \, dt = 1,\ x(0) = x(1) = 0,\ y(0) = y(1) = 0.$$

Here, x and y are arcs with values in $\mathbb{R}^n$, the Lagrangian Λ is continuously differentiable and convex, and Λ satisfies, for a certain positive constant κ,

$$\kappa |(v, w)|^2 \leqslant \Lambda(v, w) \ \ \forall (v, w) \in \mathbb{R}^n \times \mathbb{R}^n.$$

For the purposes of finding the periodic trajectory, it is required to prove the existence of a solution, and to assert the necessary conditions (that is, an appropriate multiplier rule).

The existence of a (unique) solution (x_*, y_*) is a straightforward application of the direct method. We focus here on the issue of writing the necessary conditions.

In this context, it would be cheating (or shall we say, highly inappropriate) to simply *assume* that (x_*, y_*) is either piecewise smooth or Lipschitz. We require a multiplier rule that makes no such prior assumption. Thus, we seek to apply Theorem 17.9, for the purposes of which we need to check that its growth hypothesis holds. But in the present context, we have

$$D_{x,y} \Lambda = (0,0), \ \ D_{x,y} \psi(x, y, x_*'(t), y_*'(t)) = (0, x_*'(t)),$$

where $\psi(x, y, v, w)$ is given by $y \bullet v$. We may therefore take $c = 0$ and $d(t) = |x_*'(t)|$ in order to verify the required hypothesis. This allows us to invoke Theorem 17.9, and the analysis proceeds from there; see Exer. 21.30 for the details. □

17.11 Exercise. In Exer. 16.12, one shows that the problem

$$\min \int_0^\pi x'(t)^2 dt \ : \ x \in \mathrm{AC}[0, \pi], \ \int_0^\pi x(t)^2 dt = \pi/2, \ x(0) = x(\pi) = 0$$

admits a solution. Prove that it lies in $C^\infty[0, \pi]$, so that it is, in fact, the solution of the problem examined earlier in Exer. 14.23. □

17.12 Exercise. (The Sturm-Liouville problem)

The following type of boundary-value problem arises in a variety of applications in partial differential equations and physics, notably in elasticity and acoustics:

$$-\frac{d}{dt}\{P(t)\,u'(t)\}+Q(t)\,u(t) = \lambda\,R(t)\,u(t),\;\; u \in C^2[0,1],\;\; u(0) = u(1) = 0. \quad (*)$$

Here, P, Q, and R are given continuously differentiable functions, with P and R assumed positive; $\lambda \in \mathbb{R}$ is an undetermined parameter. Evidently, the zero function is a solution of $(*)$, for any λ. It turns out that only for certain (countably many) values of λ will the problem $(*)$ admit a nontrivial solution (that is, one which is not identically zero). Such a λ is called an *eigenvalue*.

A complete variational characterization of the eigenvalues can be obtained. We content ourselves here with proving that there is a minimal eigenvalue λ_*, and that it can be characterized with the help of the following problem in the calculus of variations:[1]

$$\text{minimize } \; I(x) := \int_0^1 \{P(t)\,x'(t)^2+Q(t)\,x(t)^2\}\,dt \;\text{ subject to}$$

$$x \in \mathrm{AC}[0,1],\;\; H(x) := \int_0^1 R(t)\,x(t)^2\,dt = 1,\;\; x(0) = x(1) = 0. \quad (**)$$

(a) Prove that a solution x_* of the problem $(**)$ exists.

(b) Apply necessary conditions in order to deduce that x_* lies in $C^2[0,1]$, and that the function $u = x_*$ satisfies $(*)$ for some $\lambda_* \in \mathbb{R}$.

(c) Let λ be any eigenvalue. Show that there is an associated nontrivial solution u of $(*)$ satisfying

$$\int_0^1 R(t)\,u(t)^2\,dt = 1.$$

Deduce that $\lambda = I(u) \geqslant I(x_*)$.

(d) Prove that

$$\lambda_* = I(x_*) = \min\left\{\frac{I(x)}{H(x)} : x \in C^2[0,1],\;\; x(0) = x(1) = 0,\;\; H(x) > 0\right\}$$

is the minimal eigenvalue of the boundary-value problem $(*)$.

The functional $I(x)/H(x)$ appearing here is known as the *Rayleigh quotient*; the reader will discern a relationship between this result and Exer. 13.1. □

[1] The analysis will be more complete than some, however, in that no tacit assumption is made concerning the existence and the regularity of the solution to the variational problem.

Chapter 18
Nonsmooth Lagrangians

We now take a step in a direction never envisaged by the classical theory: the introduction of nonsmooth Lagrangians (as opposed to nonsmooth solutions). This is a modern issue that stems from new applications in such disciplines as engineering, economics, mechanics, and operations research. At the same time, this factor will play an essential role in the proof of such theorems as the multiplier rules of the preceding chapter, even when the underlying problems have smooth data.

18.1 The Lipschitz problem of Bolza

We consider a version of what is known in the calculus of variations as the *problem of Bolza*: the minimization of the (Bolza) functional

$$J(x) \;=\; \ell_0\big(x(a)\big) + \ell_1\big(x(b)\big) + \int_a^b \Lambda\big(t, x(t), x'(t)\big)\,dt$$

over all arcs $x : [\,a,b\,] \to \mathbb{R}^n$ satisfying the constraints

$$x(a) \in C_0\,,\;\; x(b) \in C_1\,,\;\; x'(t) \in V(t) \;\text{ a.e.}$$

where $[\,a,b\,]$ is a given fixed interval in $\mathbb{R}$. The reader will recall that an arc x is said to be admissible for the problem if x satisfies the given constraints, and if the integral in the cost $J(x)$ above is well defined and finite.

We are given an admissible arc x_* which is a strong local minimizer: for some $\varepsilon > 0$, for any admissible arc x satisfying $\|\,x - x_*\| \leqslant \varepsilon$, we have $J(x_*) \leqslant J(x)$. We proceed to state the main result of the chapter, beginning with its hypotheses.

Hypotheses. The Lagrangian $\Lambda(t,x,v)$, a mapping from $[\,a,b\,] \times \mathbb{R}^n \times \mathbb{R}^n$ to $\mathbb{R}$, is measurable with respect to t and Lipschitz with respect to (x,v) near x_* in the fol-

F. Clarke, *Functional Analysis, Calculus of Variations and Optimal Control*, Graduate Texts in Mathematics 264, DOI 10.1007/978-1-4471-4820-3_18,

lowing sense: for a summable function $k : [a,b] \to \mathbb{R}$, we have, for almost all t, for all $x, y \in B(x_*(t), \varepsilon)$ and $v, w \in V(t)$,

$$\left|\Lambda(t,x,v) - \Lambda(t,y,w)\right| \leqslant k(t)\{\,|x-y| + |v-w|\,\}. \tag{LH}$$

(Observe that differentiability of Λ is not assumed.) The other assumptions on the data are as follows: C_0, C_1 are closed subsets of $\mathbb{R}^n$, the real-valued functions ℓ_0, ℓ_1 are locally Lipschitz on $\mathbb{R}^n$, and V is a measurable mapping from $[a,b]$ to the closed convex subsets of $\mathbb{R}^n$.

Finally, we require the following **Interiority Hypothesis**: there is a positive δ such that

$$B\big(x_*'(t), \delta\big) \subset V(t) \text{ a.e.}$$

We observe that the Lipschitz hypothesis (LH) above evidently holds if V is uniformly bounded and Λ is locally Lipschitz. Thus, the following theorem applies in particular to a *weak* local minimum x_* in the class $\mathrm{Lip}[a,b]$, and for a locally Lipschitz Lagrangian Λ, by taking $V(t)$ to be of the form $B(x_*'(t), \delta)$.

18.1 Theorem. *Under the above hypotheses, there exists an arc p which satisfies the* **Euler inclusion:**

$$p'(t) \in \operatorname{co}\big\{\,\omega : \big(\omega, p(t)\big) \in \partial_L \Lambda\big(t, x_*(t), x_*'(t)\big)\big\} \text{ a.e. } t \in [a,b] \tag{E}$$

together with the **Weierstrass condition:** *for almost every t,*

$$\Lambda\big(t, x_*(t), v\big) - \Lambda\big(t, x_*(t), x_*'(t)\big) \geqslant \langle p(t), v - x_*'(t)\rangle \;\; \forall v \in V(t) \tag{W}$$

and the **transversality condition:**

$$p(a) \in \partial_L \ell_0\big(x_*(a)\big) + N^L_{C_0}\big(x_*(a)\big),\; -p(b) \in \partial_L \ell_1\big(x_*(b)\big) + N^L_{C_1}\big(x_*(b)\big). \tag{T}$$

In this theorem statement, the reader will note that the Weierstrass condition has a familiar look. The transversality condition is rather more complicated than the ones we have seen, but is along the same lines. The Euler inclusion (E), however, is truly novel in appearance; let us examine it more closely.

In writing it, by convention, the limiting subdifferential $\partial_L \Lambda$ (Def. 11.10) is taken with respect to the (x, v) variables for each fixed t. As we know, this reduces to a singleton (the gradient) when Λ is continuously differentiable in (x, v). Thus, in that case, (E) is equivalent to

$$p'(t) = \nabla_x \Lambda\big(t, x_*(t), x_*'(t)\big),\;\; p(t) = \nabla_v \Lambda\big(t, x_*(t), x_*'(t)\big).$$

This is the familiar integral form of the Euler equation expressed in terms of the costate p.

The examples that follow suggest why it is useful to weaken the regularity hypotheses on the Lagrangian; other such instances will be encountered later.

18.2 Example. Consider the problem of finding the path of least time between the points $(0,0)$ and $(2,2)$, where the speed of travel is v_0 (constant) in the left halfspace $t < 1$, and v_1 in the right halfspace $t > 1$. The problem may be cast as follows:

$$\text{minimize} \quad \int_0^2 \mu(t)\sqrt{1+x'(t)^2}\,dt\,,$$

where

$$\mu(t) = \begin{cases} v_0^{-1} & \text{if } t < 1 \\ v_1^{-1} & \text{if } t > 1. \end{cases}$$

Note that the Lagrangian is discontinuous in its t variable. Suppose now that x_* in $\text{Lip}[0,2]$ is a solution of the problem. Limiting attention to $[0,1]$, we see that the restriction of x_* to $[0,1]$ solves a version of the basic problem in which the Lagrangian is $(1+v^2)^{1/2}$. It follows (as we have seen) that it is an affine function. The same conclusion holds on $[1,2]$. Thus, x_* is piecewise affine on $[0,2]$, with a possible corner at $t = 1$.

Because the costate p is continuous, the necessary conditions of Theorem 18.1 go beyond this in implying

$$p(1) = \frac{\mu(1-)\,x_*'(1-)}{\sqrt{1+x_*'(1-)^2}} = \frac{\mu(1+)\,x_*'(1+)}{\sqrt{1+x_*'(1+)^2}}\,,$$

which, together with the boundary conditions, uniquely specifies the piecewise affine function. In a more geometric vein, let us denote by θ_0 and θ_1 the angles of incidence of the path with the line $t = 1$. Then the relation found above is equivalent to

$$\frac{\sin\theta_0}{v_0} = \frac{\sin\theta_1}{v_1}\,,$$

which we recognize as Snell's law of refraction. This is not a coincidence, of course: *Fermat's principle* in optics asserts that light rays propagating in inhomogeneous media follow paths that minimize time. □

18.3 Example. We turn now to an example involving friction in mechanics. The classical differential equation governing the free oscillation of the pendulum in the plane is, as we have seen (Example 14.6):

$$m\ell\,\theta''(t) + mg\sin\theta(t) = 0.$$

When an external constant force f (tangentially applied to the mass) is present, the governing equation becomes

$$m\ell\,\theta''(t) + mg\sin\theta(t) = f.$$

This is still the Euler equation for the action, now defined, however, as follows:

$$\int_{t_1}^{t_2} \left\{ \tfrac{1}{2} m\big(\ell\,\theta'(t)\big)^2 - mg\ell\big(1-\cos\theta(t)\big) + f\,\ell\,\theta(t) \right\} dt\,.$$

Suppose now that the pendulum is subject to a frictional force, one that equals -1 when $\theta > 0$ and $+1$ when $\theta < 0$ (thus, it is a force opposing movement in any direction from equilibrium). Then the governing equation becomes

$$m\ell\,\theta''(t) + mg\sin\theta(t) = \begin{cases} +1 & \text{if } \theta < 0 \\ -1 & \text{if } \theta > 0. \end{cases}$$

This "discontinuous differential equation" corresponds to the action functional

$$\int_{t_1}^{t_2} \left\{ \tfrac{1}{2} m\big(\ell\,\theta'(t)\big)^2 - mg\ell\big(1-\cos\theta(t)\big) + \ell\,|\,\theta(t)| \right\} dt\,.$$

Note that this integrand is nondifferentiable with respect to the state θ. The extremals of this action functional, in the sense of Theorem 18.1, satisfy the discontinuous differential equation. The Euler inclusion further implies that, for a minimizing function, θ' is absolutely continuous, since (E) asserts in part that $p(t) = m\ell^2\,\theta'(t)$. The Euler inclusion also leads to a natural interpretation of the "equation" at $\theta = 0$:

$$m\ell\,\theta''(t) + mg\sin\theta(t) \in [-1,+1] \text{ if } \theta = 0. \qquad \square$$

18.4 Exercise. Show that the arc p of Theorem 18.1 satisfies

$$\big(p'(t), p(t)\big) \in \partial_C \Lambda\big(t, x_*(t), x_*'(t)\big) \text{ a.e.}$$

(This form of the Euler inclusion is often easier to calculate in practice.) $\square$

The transversality condition. Consider the problem of Bolza in the case in which ℓ_1 is smooth, C_0 is a singleton, and $C_1 = \mathbb{R}^n$; thus, $x(a)$ is prescribed, and $x(b)$ is unconstrained. Then the first part of (T) gives no information, since the normal cone to a singleton is the whole space. The second part affirms $-p(b) = \nabla\ell_1(x_*(b))$, since the normal cone to the whole space is $\{0\}$. We recover, therefore, the classical transversality condition encountered earlier in Theorem 14.19.

Evidently, the formulation of the boundary costs and constraints is more general in the problem of Bolza than it was before. Theorem 18.1 expresses the appropriate transversality conditions, as regards the endpoint constraint sets C_0 and C_1, entirely in geometric form, with normal vectors. These sets are not restricted to being classical manifolds (with or without boundary). When they are defined by functional equalities and inequalities, however, we can deduce alternate formulations of transversality that are stated in terms of multipliers, by means of characterizations such as Theorem 11.38.

Regularity. The following exercise shows that the regularity theorem 15.5 carries over to locally Lipschitz Lagrangians. Thus, solutions can inherit more regularity than the Lagrangian has to bequeath.

18.5 Exercise. In addition to the hypotheses of Theorem 18.1, suppose that x_* is Lipschitz, and that, for almost every t, the function $v \mapsto \Lambda(t, x_*(t), v)$ is strictly convex. Prove that x_* lies in $C^1[a,b]$. □

18.2 Proof of Theorem 18.1

When the Lagrangian Λ is nondifferentiable, entirely new methods must be found for deriving necessary conditions, as a look back to the classical theory will confirm. A complex of ideas based upon nonsmooth calculus, penalization, and inf-convolution makes its first appearance in the proof below.

The following technical result will be needed (and used again later). It concerns a function $f(t,x,v)$ of three variables whose limiting subdifferential with respect to the (x,v) variables is denoted by $\partial_L f$.

18.6 Proposition. *Let $f : [a,b] \times \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ be measurable in t and locally Lipschitz in (x,v), with*

$$\partial_L f(t,x,v) \subset k(t)B \;\; \forall (t,x,v),$$

where k is summable. Let q_i, p_i, x_i, and v_i be sequences of measurable functions with values in $\mathbb{R}^n$ such that

$$q_i \to q \text{ weakly in } L^1(a,b),\; p_i(t) \to p(t) \text{ a.e.},\; x_i(t) \to x(t) \text{ a.e.},\; v_i(t) \to v(t) \text{ a.e.}$$

and satisfying $|q_i(t)| \leqslant k(t)$ a.e. for each i. For certain constants $c \leqslant d$, suppose that we have

$$q_i(t) \in \operatorname{co}\{\omega : (\omega, p_i(t)) \in [c,d]\,\partial_L f(t, x_i(t), v_i(t)) + \varepsilon_i B\},\; t \in \Omega_i \text{ a.e.},$$

where $\varepsilon_i \downarrow 0$, and where Ω_i is a sequence of measurable subsets of $[a,b]$ such that $\operatorname{meas}(\Omega_i) \to b-a$. Then we have in the limit

$$q(t) \in \operatorname{co}\{\omega : (\omega, p(t)) \in [c,d]\,\partial_L f(t, x(t), v(t))\} \text{ a.e.}$$

(The notation $[c,d]\,\partial_L f$ refers here to $\{r\zeta : r \in [c,d],\, \zeta \in \partial_L f\}$.)

Proof. Let us define the multifunction Γ as follows:

$$\Gamma(t,x,v,p,\varepsilon) = \operatorname{co}\{\omega : (\omega, p) \in [c,d]\,\partial_L f(t,x,v) + |\varepsilon| B\}.$$

Then we have

$$q_i(t) \in \Gamma\big(t, x_i(t), v_i(t), p_i(t), \varepsilon_i\big), \ \ t \in \Omega_i \text{ a.e.}$$

If the weak closure theorem 6.39 can be invoked, with $Q = [a,b] \times \mathbb{R}^{3n+1}$ and $r_i = 0$, then it gives precisely the desired conclusion. We proceed to verify that the theorem applies, beginning with the hypothesis that $\Gamma(t,\cdot)$ has closed graph.

Let $(x_j, v_j, p_j, \varepsilon_j, q_j)$ be a sequence of points in the graph of $\Gamma(t,\cdot)$ converging to (x,v,p,ε,q). Then (by Carathéodory's theorem 2.6) each q_j is expressible in the form

$$q_j = \sum_{m=0}^{n} \lambda_j^m \, \omega_j^m,$$

where λ_j^m $(m = (0,1)\ldots,n)$ are the coefficients of a convex combination and ω_j^m satisfies, for each $m \in \{(0,1),\ldots,n\}$,

$$(\omega_j^m, p_j) \in r_j^m \, \zeta_j^m + \varepsilon_j B, \text{ for some } r_j^m \in [c,d], \ \zeta_j^m \in \partial_L f(t, x_j, v_j).$$

Because the multifunction $(x,v) \mapsto \partial_L f(t,x,v)$ is bounded, we may suppose, by taking subsequences, that all data sequences converge as $j \to \infty$. Because the multifunction $(x,v) \mapsto \partial_L f(t,x,v)$ has closed graph, we obtain in the limit

$$q = \sum_{m=0}^{n} \lambda^m \, \omega^m,$$

where λ^m $(m = (0,1)\ldots,n)$ are the coefficients of a convex combination and each ω^m satisfies

$$(\omega^m, p) \in r^m \, \zeta^m + \varepsilon B, \text{ for some } r^m \in [c,d], \ \zeta^m \in \partial_L f(t,x,v).$$

It follows that (x,v,p,ε,q) lies in the graph of $\Gamma(t,\cdot)$, as required.

Hypothesis (c) of Theorem 6.39 holds as a direct result of the assumptions made. We verify now the measurability hypothesis (b) of the theorem. This requires that, for any measurable functions

$$x(t)\,,\ v(t)\,,\ p(t)\,,\ \varepsilon(t)\,,$$

and for any point $z \in \mathbb{R}^n$, the function

$$t \mapsto \max\big\{ \langle z, \omega \rangle : \big(\omega, p(t)\big) \in [c,d]\, \partial_L f\big(t, x(t), v(t)\big) + |\varepsilon(t)| B \big\}$$

be measurable. But this function may be expressed as

$$\max\big\{ \langle (z,0), (\omega, p(t)) \rangle : \big(\omega, p(t)\big) \in [c,d]\, \partial_L f\big(t, x(t), v(t)\big) + |\varepsilon(t)| B \big\},$$

so that its measurability would follow (by Prop. 6.29) from the measurability of the multifunction

$$t \mapsto \big\{ \big(\omega, p(t)\big) : \big(\omega, p(t)\big) \in [c,d]\, \partial_L f\big(t, x(t), v(t)\big) + |\varepsilon(t)| B \big\}.$$

But this is the intersection of the two multifunctions

$$\Gamma_1(t) = [c,d]\,\partial_L f\big(t,x(t),v(t)\big) + |\varepsilon(t)|B, \quad \Gamma_2(t) = \mathbb{R}^n \times \{p(t)\},$$

and each of these is easily seen to be measurable (using Exer. 13.24 and Exer. 6.24 for the first, and Exer. 6.21 for the second). The required measurability therefore follows from Cor. 6.26. □

Reductive hypotheses. Before beginning the actual proof of Theorem 18.1, we identify certain additional hypotheses that can be added "for free" (without any loss of generality) by simple reformulations. We claim first that the theorem's hypotheses and conclusions are unaffected if we redefine

$$\Lambda(t,x,v) := \Lambda\big(t, \pi_t^1(x), \pi_t^2(v)\big),$$

where $\pi_t^1(x)$ denotes the projection of x onto the set $B(x_*(t),\varepsilon)$ and $\pi_t^2(v)$ is the projection of v onto $V(t)$. The optimality of x_* in the stated sense is clearly preserved by this modification, as are the required conclusions. As regards the measurability in t of the new data, note that we have (for example):

$$\big\{\pi_t^2(v)\big\} = \big\{w \in V(t) : d_{V(t)}(v) = |v-w|\big\}.$$

It follows from Prop. 6.25 and Exer. 6.30 that $t \mapsto \pi_t^2(v)$ is measurable.

Since π_t^1 and π_t^2 are globally Lipschitz, the redefinition above allows us to suppose that the Lipschitz condition (LH) (p. 348) holds *globally*. By similar arguments, we may suppose that ℓ_0 and ℓ_1 are bounded below and globally Lipschitz, and that C_0 is compact.

It is clear that we change nothing in taking $k(t) \geqslant 1$. It is also easy to see, by a simple reformulation, that we may take $x_* \equiv 0$ and $[a,b] = [0,1]$. The final reduction that we make is the following:

18.7 Proposition. *It suffices to prove the theorem when V is replaced by*

$$V_\eta(t) := \big\{v \in V(t) \cap B\big(0,1/\eta\big) : v + \eta B \subset V(t)\big\},$$

for any $\eta > 0$ sufficiently small.

Proof. Note that $x_* = 0$ continues to be admissible and optimal for the problem when V is replaced by V_η, for $\eta < \delta$. If the reduced theorem is proved as stated, then there is a sequence of arcs p_i satisfying the conclusions of the theorem for V replaced by V_{η_i}, where $\eta_i \downarrow 0$. It follows from the Euler inclusion that

$$|(p_i'(t), p_i(t)| \leqslant k(t) \text{ a.e.}$$

This implies (exercise) $\min_{[0,1]} |p_i(t)|$ is bounded above independently of i. It then follows that p_i is bounded in $C[0,1]$, as well as equicontinuous. By Ascoli's theorem, and by weak compactness in $L^1[0,1]$ (see Prop. 6.17), taking subsequences (without relabeling), we find an arc p such that $p_i \to p'$ weakly and

$p_i(t) \to p(t)\ \forall t$. It is clear that the arc p continues to satisfy the transversality condition (T), and easy to see that it satisfies (W) for $V(t)$. That p satisfies the Euler inclusion (E) follows from Prop. 18.6, with $f = \Lambda,\ c = d = 1$. □

We now prove the theorem (a fairly arduous task with many implicit exercises), under the reductive hypotheses above, with V replaced by V_η $(\eta < \delta)$, and under an additional hypothesis whose removal will constitute the last step in the proof.

Temporary hypothesis:

(TH) $C_1 = \mathbb{R}^n$ (so that there is no explicit constraint on $x(1)$).

A. We proceed to define via penalization a sequence of *decoupled* problems converging in an appropriate sense to (P). (The reader will come to understand this cryptic phrase, in time.) We introduce, for a given sequence of positive numbers n_i tending to ∞, the functions

$$\ell_i^1(y) = \min_{\beta \in \mathbb{R}^n} \{ \ell_1(\beta) + n_i |y - \beta|^2 \}. \tag{1}$$

These quadratic inf-convolutions, the Moreau-Yosida approximations to ℓ_1, have been encountered before in §7.4. Since ℓ_1 is globally Lipschitz (of rank K_1, say), we calculate

$$\ell_i^1(y) \geqslant \min_{\beta \in \mathbb{R}^n} \{ \ell_1(y) - K_1 |y - \beta| + n_i |y - \beta|^2 \} = \ell_1(y) - K_1^2/(4n_i).$$

Thus there is a constant c independent of i such that

$$\ell_i^1 \leqslant \ell_1 \leqslant \ell_i^1 + c/n_i .$$

We set

$$\Lambda_i(t,x,v) = \min_{u \in \mathbb{R}^n} \{ \Lambda(t,u,v) + n_i k(t) |u - x|^2 \}.$$

(This decouples x and v.) The global Lipschitz condition (LH) implies that the minimum is attained. Because $\Lambda(t,u,v)$ is continuous in u, the minimum over u is the same as the infimum for u with rational coordinates. Then Λ_i is the countable infimum of functions which are LB measurable, since, for each u, the function

$$(t,x,v) \mapsto \Lambda(t,u,v) + n_i k(t) |u - x|^2$$

is measurable in t and continuous in (x,v). Thus Λ_i is itself LB measurable, and the following functional is well defined:

$$J_i(x) := \ell_0\big(x(0)\big) + \ell_i^1\big(x(1)\big) + \int_0^1 \Lambda_i\big(t, x(t), x'(t)\big)\, dt .$$

We define I_i to be the infimum of $J_i(x)$ over all arcs x satisfying

$$x(0) \in C_0\,,\;\; x'(t) \in V_\eta(t) \text{ a.e.},\;\; |x(0)| \leqslant \varepsilon/2\,,\;\; \int_0^1 |x'(t)|\,dt \;\leqslant \varepsilon/2\,. \tag{2}$$

Note that these constraints imply $|x(t)| \leqslant \varepsilon$. Because V_η is bounded, we have (for some constant c_0)

$$c_0 \leqslant I_i \leqslant J_i(0) \leqslant \ell_0(0)+\ell_1(0)+\int_0^1 \Lambda(t,0,0)\,dt \;=\; J(0)\,.$$

Let y_i be an arc satisfying the constraints (2) together with $J_i(y_i) \leqslant I_i + n_i^{-1}$.

Lemma 1. *There is a measurable function w_i such that $w_i(t)$ is (almost everywhere) a point at which the minimum defining $\Lambda_i(t, y_i(t), y_i'(t))$ is achieved*:

$$\Lambda_i\big(t, y_i(t), y_i'(t)\big) \;=\; \Lambda\big(t, w_i(t), y_i'(t)\big) + n_i k(t)|w_i(t) - y_i(t)|^2 \text{ a.e.}$$

The proof of the lemma uses the multifunction Γ defined by

$$\Gamma(t) \;=\; \big\{ w \in \mathbb{R}^n : \Lambda_i\big(t, y_i(t), y_i'(t)\big) \;=\; \Lambda\big(t, w, y_i'(t)\big) + n_i k(t)|w - y_i(t)|^2 \big\}.$$

Note that, for almost every t, the set $\Gamma(t)$ is nonempty (since the minimum defining Λ_i is attained) and closed (by the continuity in w). As pointed out previously, the function Λ_i is LB measurable, so that the map $t \mapsto \Lambda_i(t, y_i(t), y_i'(t))$ is measurable. It follows that, for a certain function $g(t,w)$ that is measurable in t and continuous in w, we have

$$\Gamma(t) \;=\; \big\{ w \in \mathbb{R}^n \;:\; g(t,w) \;=\; 0 \big\}.$$

Then Γ is closed-valued and measurable (Prop. 6.25) and therefore admits a measurable selection (Cor. 6.23), as asserted by Lemma 1.

Recall that $x_* = 0$. It is in the following sense that the sequence of decoupled problems converges to the original one:

Lemma 2. $\lim_{i\to\infty} I_i \;=\; J(0)\,.$

To see this, let w_i be the function provided by Lemma 1. The equality defining it, together with the Lipschitz hypothesis (LH), gives rise to the following estimate: $|w_i(t) - y_i(t)| \leqslant n_i^{-1}$ a.e. With the help of this, we now calculate

$$\begin{aligned}
I_i + n_i^{-1} &\geqslant J_i(y_i) \\
&= \ell_0\big(y_i(0)\big) + \ell_i^1\big(y_i(1)\big) + \int_0^1 \big\{ \Lambda\big(t, w_i, y_i'\big) + n_i k(t)|w_i - y_i|^2 \big\}\,dt \\
&\geqslant \ell_0\big(y_i(0)\big) + \ell_1\big(y_i(1)\big) - c/n_i + \int_0^1 \big\{ \Lambda\big(t, y_i, y_i'\big) - k(t)|w_i - y_i| \big\}\,dt \\
&\geqslant J(0) - c/n_i - \|k\|_1/n_i\,,
\end{aligned}$$

and the assertion of Lemma 2 follows.

We may view the problem defining I_i as one that is defined relative to the couples $(x(0), x')$ in the complete metric space $\mathbb{R}^n \times L^1$ lying in the closed set S defined by (2). The lemma implies that the couple $(0,0)$ is ε_i^2-optimal for the problem (that is, gives a value to the cost that is within ε_i^2 of the infimum), for some positive sequence ε_i tending to 0. We apply Theorem 5.19 to deduce the existence of an element $(x_i(0), x_i') \in S$ satisfying

$$|x_i(0)| + \int_0^1 \big|x_i'(t)\big|\,dt \leqslant \varepsilon_i, \tag{3}$$

and which minimizes over S the functional $\widetilde{J}_i$ defined by

$$\widetilde{J}_i\big(x(0), x'\big) = \ell_0\big(x(0)\big) + \varepsilon_i|x(0) - x_i(0)| + \ell_i^1\big(x(1)\big) \\ + \int_0^1 \min_u \big\{ \Lambda\big(t, u, x'(t)\big) + n_i k(t)|u - x(t)|^2 + \varepsilon_i|x'(t) - x_i'(t)| \big\}\,dt.$$

We pass to a subsequence (without relabeling) so as to have $x_i'(t) \to 0$ a.e.

B. We now fix i and reformulate the optimality of x_i for $\widetilde{J}_i$ in a more useful manner, one that will allow us to identify an arc p_i that is "close" to satisfying the necessary conditions. Let u_i be a measurable function such that, almost everywhere, the minimum

$$\min_u \big\{ \Lambda\big(t, u, x_i'(t)\big) + n_i k(t)|u - x_i(t)|^2 \big\}$$

is achieved at $u_i(t)$; as in Lemma 2, we have $|u_i(t) - x_i(t)| \leqslant n_i^{-1}$ a.e., which implies that $u_i(t) \to 0$ a.e.

Now let β_i be a point achieving the minimum in (1) when $y = x_i(1)$. It follows readily that $\beta_i \to x_*(1) = 0$. We proceed to define an arc p_i via

$$p_i'(t) = 2n_i k(t)\big(x_i(t) - u_i(t)\big), \quad p_i(1) = -2n_i\big(x_i(1) - \beta_i\big).$$

Then we have (by choice of β_i)

$$-p_i(1) \in \partial_P \ell_1\big(\beta_i\big). \tag{4}$$

Observe that, from the way $\ell_i^1(y)$ is defined, we have

$$\ell_i^1(y) \leqslant \ell_1\big(\beta_i\big) + n_i|y - \beta_i|^2 \quad \forall y,$$

with equality for $y = x_i(1)$. We shall also need below the identity

$$n_i k|u - x|^2 = n_i k|u_i - x_i|^2 - \langle p_i', u - u_i\rangle + \langle p_i', x - x_i\rangle + n_i k\big|(x - x_i) - (u - u_i)\big|^2.$$

Define the cost functional $\Phi(u, \alpha, v)$ on $L^\infty \times \mathbb{R}^n \times L^1$ by

$$\ell_0(\alpha) + \varepsilon_i |\alpha - x_i(0)| - \langle p_i(0), \alpha \rangle + n_i |x(1) - \beta_i|^2 + \langle p_i(1), x(1) \rangle$$
$$+ \int_0^1 \big\{ \Lambda\big(t, u(t), v(t)\big) - \langle p_i(t), v(t) \rangle - \langle p_i'(t), u(t) \rangle + \varepsilon_i |v(t) - x_i'(t)| \big\}\, dt$$
$$+ 2n_i \int_0^1 k(t) \big\{ |u(t) - u_i(t)|^2 + |x(t) - x_i(t)|^2 \big\}\, dt,$$

where

$$x(t) = \alpha + \int_0^t v(s)\, ds.$$

Then an elementary calculation using the observation and identity above, together with integration by parts, shows that the optimality of $(x_i(0), x_i')$ for $\widetilde{J}_i$ translates into the following: for a certain constant c_i, we have

$$\Phi(u, \alpha, v) \geqslant \widetilde{J}_i\big(x_i(0), x_i'\big) + c_i$$

whenever (u, α, v) satisfies

$$\alpha \in C_0,\ |\alpha| \leqslant \varepsilon/2,\ v(t) \in V_\eta(t) \text{ a.e.},\ \int_0^1 |v(t)|\, dt \leqslant \varepsilon/2, \tag{5}$$

with equality when $(u, \alpha, v) = (u_i, x_i(0), x_i')$. Thus, $\Phi(u, \alpha, v)$ is minimized relative to the constraints (5) at $(u_i, x_i(0), x_i')$.

It is easy to see (by substituting for x and p) that the last two boundary terms in the expression for Φ may be rewritten in the form

$$n_i \big\{ |x(1) - x_i(1)|^2 + |\beta_i|^2 - |x_i(1)|^2 \big\}.$$

It follows then that the modified functional $\Psi(u, \alpha, v)$ defined by

$$\ell_0(\alpha) + \varepsilon_i |\alpha - x_i(0)| - \langle p_i(0), \alpha \rangle + n_i |x(1) - x_i(1)|^2$$
$$+ \int_0^1 \big\{ \Lambda(t, u(t), v(t)) - \langle p_i(t), v(t) \rangle - \langle p_i'(t), u(t) \rangle + \varepsilon_i |v(t) - x_i'(t)| \big\}\, dt$$
$$+ 2n_i \int_0^1 k(t) \big\{ |u(t) - u_i(t)|^2 + |x(t) - x_i(t)|^2 \big\}\, dt$$

is minimized relative to the constraints (5) at $(u_i, x_i(0), x_i')$.

C. The next step consists of a variational analysis (for i still fixed) of the minimum of Ψ just mentioned. Let us first fix $u = u_i$ and $v = x_i'$. Then the function $\alpha \mapsto \Psi(u_i, \alpha, x_i')$ attains a local minimum (for i sufficiently large, since $x_i(0) \to 0$) relative to $\alpha \in C_0$ at $x_i(0)$. The corresponding necessary condition is

$$p_i(0) \in \partial_L \big\{ \ell_0 + I_{C_0} \big\}\big(x_i(0)\big) + \varepsilon_i B. \tag{6}$$

This, together with (4), is the precursor of the transversality condition (T).

We now exploit the minimum in v of $\Psi(u_i, x_i(0), v)$ to derive a forerunner of the Weierstrass condition. The constraint on v in (5) is slack for i sufficiently large, and a simple argument by contradiction shows that we have, for almost every t,

$$\begin{aligned}\langle p_i(t), v\rangle - \Lambda\big(t, u_i(t), v\big) - \varepsilon_i |v - x_i'(t)| \leqslant \\ \langle p_i(t), x_i'(t)\rangle - \Lambda\big(t, u_i(t), x_i'(t)\big) \;\; \forall v \in V_\eta(t). \quad (7)\end{aligned}$$

Let us give this argument. If (7) does not hold, then there exists $r > 0$ and a subset Σ of $[a,b]$ of positive measure m such that, for some measurable function w defined on Σ and taking values in $V_\eta(t)$, we have

$$\begin{aligned}\Lambda(t, u_i(t), w(t)) + \varepsilon_i |w(t) - x_i'(t)| - \langle p_i(t), w(t)\rangle \leqslant \\ \Lambda(t, u_i(t), x_i'(t)) - \langle p_i(t), x_i'(t)\rangle - r, \;\; t \in \Sigma \text{ a.e.}\end{aligned}$$

Of course, m can be taken arbitrarily small. If we let v be the function equal to w on Σ and equal to x_i' elsewhere, and if $x(t)$ signifies

$$x_i(0) + \int_0^t v(s)\, ds,$$

then we have $\|x - x_i\| \leqslant Km$ for a constant K independent of m (the boundedness of V_η is used for this). It follows that for m sufficiently small we have

$$\Psi\big(u_i, x_i(0), v\big) - \Psi\big(u_i, x_i(0), x_i'\big) \leqslant -rm + K^2 n_i \big(1 + 2\|k\|_1\big) m^2 < 0.$$

Furthermore, v satisfies the constraint in (5) if m is small enough. This contradicts the optimality of $(u_i, x_i(0), x_i')$ and concludes the argument.

Making use of the evident estimate

$$|x(t) - x_i(t)|^2 \leqslant 2|x(0) - x_i(0)|^2 + 2\int_0^1 |x'(s) - x_i'(s)|^2\, ds$$

and rearranging, we deduce that the cost functional $\Psi^+(u, \alpha, v)$ defined by

$$\begin{aligned}&\ell_0(\alpha) + \varepsilon_i \big|\alpha - x_i(0)\big| - \langle p_i(0), \alpha\rangle + 6n_i |\alpha - x_i(0)|^2 \\ &+ \int_0^1 \big\{\Lambda\big(t, u(t), v(t)\big) - \langle p_i(t), v(t)\rangle - \langle p_i'(t), u(t)\rangle + \varepsilon_i |v(t) - x_i'(t)|\big\}\, dt \\ &+ 2n_i \int_0^1 k(t)|u(t) - u_i(t)|^2\, dt + 6n_i \int_0^1 k(t)|v(t) - x_i'(t)|^2\, dt\end{aligned}$$

also attains a minimum relative to the constraints (5) at $(u_i, x_i(0), x_i')$.

Setting $\alpha = x_i(0)$ and $v = x_i'$ in Ψ^+, the attainment of the minimum relative to $u \in L^\infty$ implies (by measurable selections) that for t a.e., it is the case that $u_i(t)$ minimizes freely the integrand in Ψ^+. This fact yields

$$p_i'(t) \in \partial_P \{ \Lambda(t, \cdot, x_i'(t)) \} (u_i(t)) \text{ a.e.},$$

which in turn gives

$$|p_i'(t)| \leqslant k(t) \text{ a.e.} \tag{8}$$

When the constraint on v in (5) is slack, it follows (Theorem 6.32) that for almost every t, the minimum with respect to $(u,v) \in \mathbb{R}^n \times V_\eta(t)$ of the integrand in Ψ^+ is attained at $(u_i(t), x_i'(t))$; this implies an intermediate version of the Euler inclusion:

$$(p_i'(t), p_i(t)) \in \partial_L \Lambda(t, u_i(t), x_i'(t)) + \{0\} \times \varepsilon_i B, \; t \in \Omega_i \text{ a.e.} \tag{9}$$

where $\Omega_i := \{ t \in [0,1] : x_i'(t) \in \operatorname{int} V_\eta(t) \}$. (The limiting subdifferential arises upon applying the sum rule of Theorem 11.16.) One may show (exercise) that the measure of Ω_i tends to 1 as $i \to \infty$, in light of the Interiority Hypothesis.

D. The next step is to let i tend to infinity. The conditions (8) and (4) allow us to deduce (for a subsequence, without relabeling) that p_i converges uniformly to an arc p and p_i' converges weakly in $L^1(0,1)$ to p' (see Exer. 6.42). Passing to the limit in (4) and (6) (taking note of (3), and recalling that $\beta_i \to x_*(1) = 0$), we see that p satisfies the transversality condition (T) (for $C_1 = \mathbb{R}^n$). From (7) (recalling that u_i and x_i' converge almost everywhere to 0) we conclude that almost everywhere we have

$$\langle p(t), v \rangle - \Lambda(t,0,v) \leq \langle p(t), 0 \rangle - \Lambda(t,0,0) \;\; \forall v \in V_\eta(t),$$

which is the desired Weierstrass condition. Finally, the Euler inclusion (E) follows from (9), in view of Prop. 18.6 (with $c = d = 1$).

The theorem has therefore been proved, in the presence of the Temporary Hypothesis (TH).

E. We now proceed to the removal of the Temporary Hypothesis. The case of an arbitrary C_1 can be reduced to the one in which $C_1 = \mathbb{R}^n$ by an exact penalization device, as follows. A simple argument by contradiction shows that for some $K > 0$ sufficiently large, x_* solves the problem of minimizing

$$J_K(x) := \ell_0(x(0)) + \ell_1(x(1)) + K d_{C_1}(x(1)) + \int_0^1 \Lambda(t, x(t), x'(t))\, dt$$

over the arcs x satisfying

$$x(0) \in C_0, \; \|x - x_*\| \leqslant \varepsilon/2, \; x'(t) \in V_\eta(t) \text{ a.e.}$$

The argument goes as follows. If the assertion is false, then there exists for each positive integer j an arc x_j admissible for this problem with $J_j(x_j) < J_j(x_*) = J(x_*)$. Since $J_j(x_j)$ is bounded below, it follows that $d_{C_1}(x_j(1)) \to 0$. Let σ_j be a closest point in C_1 to $x_j(1)$. Then for j sufficiently large, the arc

$$y_j(t) := x_j(t) + t(\sigma_j - x_j(1))$$

satisfies

$$y_j(0) \in C_0, \;\; y_j(1) \in C_1, \;\; y_j'(t) \in V(t) \text{ a.e.}$$

(This uses the fact that $V_\eta(t) + \eta B \subset V(t)$.) A routine estimate using the Lipschitz hypothesis shows that, for a certain constant K, we have

$$J(y_j) \leqslant J(x_j) + K|\sigma_j - x_j(1)| = J(x_j) + K\, d_{C_1}\big(x_j(1)\big).$$

Then, for $j > K$, we deduce

$$J(y_j) \leqslant J(x_j) + K\, d_{C_1}\big(x_j(1)\big) \leqslant J_j(x_j) < J(x_*),$$

contradicting the optimality of x_*.

The new penalized problem above satisfies (TH), so we may apply the theorem (which has been proved under this additional hypothesis) to deduce the existence of an arc p satisfying the Euler inclusion, the Weierstrass condition (for V_η, as agreed), and the transversality condition corresponding to the new penalized cost. At $t = 0$, this is already what we wish to have. It remains to see that the correct transversality holds at $t = 1$. The arc p satisfies

$$\begin{aligned} -p(1) \in \partial_L\big(\ell_1 + K d_{C_1}\big)\big(x_*(1)\big) &\subset \partial_L \ell_1\big(x_*(1)\big) + K\,\partial_L d_{C_1}\big(x_*(1)\big) \\ &\subset \partial_L \ell_1\big(x_*(1)\big) + N^L_{C_1}\big(x_*(1)\big), \end{aligned}$$

by Theorem 11.16 and Prop. 11.34, confirming (T).

This completes the proof of Theorem 18.1.

18.3 Sufficient conditions by convexity

As the reader well knows, we generally ask of our necessary conditions that they be sufficient when the problem is convex. The following illustrates this for the problem of Bolza, and extends Theorem 15.9.

18.8 Theorem. *Let $x_* \in AC[a,b]$ be admissible for the problem of minimizing*

$$J(x) = \ell_0\big(x(a)\big) + \ell_1\big(x(b)\big) + \int_a^b \Lambda\big(t, x(t), x'(t)\big)\, dt$$

subject to

$$x(a) \in C_0, \;\; x(b) \in C_1.$$

Let the functions ℓ_0, ℓ_1 and the sets C_0, C_1 be convex, and suppose that $\Lambda(t,x,v)$ is measurable in t and convex in (x,v). If there exists an arc p satisfying the Euler inclusion (E) *and the transversality condition* (T) *of* Theorem 18.1, *then x_* is a global minimizer.*

Proof. Since $\Lambda(t,\cdot,\cdot)$ is convex, $\partial_L \Lambda$ coincides with the subdifferential of convex analysis (Prop. 11.23), and the Euler inclusion asserts that the point $(p'(t),p(t))$ belongs to $\partial\Lambda(t,x_*(t),x_*'(t))$ a.e. (We remark that the Weierstrass condition (W) follows from this.) Thus, for any arc x satisfying the constraints of the problem, we have

$$\Lambda(t,x,x') \geqslant \Lambda\big(t,x_*,x_*'\big)+(p',p)\bullet(x-x_*,x'-x_*') \text{ a.e.},$$

which implies that $J(x)$ is well defined. Since the sets C_0, C_1 and the functions ℓ_0, ℓ_1 are convex, the transversality condition yields the existence of

$$\nu_0 \in N_{C_0}\big(x_*(a)\big),\ \nu_1 \in N_{C_1}\big(x_*(b)\big),\ \zeta_0 \in \partial\ell_0\big(x_*(a)\big),\ \zeta_1 \in \partial\ell_1(x_*(b))$$

such that

$$p(a) = \zeta_0+\nu_0,\ \ -p(b) = \zeta_1+\nu_1,$$

where the normal cones and subdifferentials are understood in the sense of convex analysis. We write

$$\begin{aligned} J(x)-J(x_*) &= \ell_0\big(x(a)\big)-\ell_0\big(x_*(a)\big)+\ell_1\big(x(b)\big)-\ell_1\big(x_*(b)\big)\\ &\qquad+\int_a^b \big\{\Lambda(t,x,x')-\Lambda(t,x_*,x_*')\big\}\,dt\\ &\geqslant \langle\,\zeta_0,x(a)-x_*(a)\rangle+\langle\,\zeta_1,x(b)-x_*(b)\rangle\\ &\qquad+\int_a^b (p',p)\bullet(x-x_*,x'-x_*')\,dt\\ &= \langle\,p(a)-\nu_0,x(a)-x_*(a)\rangle-\langle\,p(b)+\nu_1,x(b)-x_*(b)\rangle\\ &\qquad+\langle\,p(b),x(b)-x_*(b)\rangle-\langle\,p(a),x(a)-x_*(a)\rangle\\ &= \langle -\nu_0,x(a)-x_*(a)\rangle-\langle\,\nu_1,x(b)-x_*(b)\rangle \geqslant 0, \end{aligned}$$

by the characterization of normal vectors to a convex set. □

18.9 Example. We had identified in Example 17.5 an arc x_* that, potentially, might be the solution of the problem considered there. Let us show how convexity can be exploited to confirm that it is so. Recall that x_* was identified via the multiplier rule: for a certain bounded function $\lambda(t) \geqslant 0$ a.e., it is an extremal of the augmented Lagrangian

$$\Lambda+\lambda\,\varphi = v^2+4\,x-\lambda(t)(v+2).$$

Because this function is convex in (x,v), it follows from Theorem 18.8 that, for any arc x satisfying the boundary conditions, we have

$$\begin{aligned} \int_0^3 \big\{x'^2+4\,x-\lambda(t)\big(x'+2\big)\big\}\,dt &\geqslant \int_0^3 \big\{x_*'^2+4\,x_*-\lambda(t)\big(x_*'+2\big)\big\}\,dt\\ &= \int_0^3 \big\{x_*'^2+4\,x_*\big\}\,dt\,, \end{aligned}$$

since, almost everywhere, $\lambda(t) = 0$ whenever $x_*'(t) + 2 \neq 0$. Suppose now that x also satisfies $x'(t) \geqslant -2$ a.e. Then, because $\lambda(t) \geqslant 0$ a.e., we deduce

$$\int_0^3 \{x'^2 + 4x\}\,dt \geqslant \int_0^3 \{x_*'^2 + 4x_*\}\,dt,$$

confirming x_* as a solution of the original problem. □

18.10 Exercise. Let the *horizon* $T > 0$ be fixed.

(a) Prove that the problem

$$\min \int_0^T \{\,|x'(t)|^2/2 + |x(t)|\,\}\,dt \,:\, x(0) = A,\ x(T) = B,\ x \in \text{Lip}[0,T]$$

has a unique solution x_*, and that $x_* \in C^1[0,T]$.

(b) Show that for $n = 1$, $A = -1$, $B = 1$, and in the case $T > 2\sqrt{2}$ (thus, for sufficiently long horizons), the solution x_* is given by:

$$x_*(t) = \begin{cases} -t^2/2 + \sqrt{2}\,t - 1 & \text{if } 0 \leqslant t \leqslant \sqrt{2} \\ 0 & \text{if } \sqrt{2} < t < T - \sqrt{2} \\ t^2/2 - (T - \sqrt{2})\,t + (T - \sqrt{2})^2/2 & \text{if } T - \sqrt{2} \leqslant t \leqslant T. \end{cases}$$

The segment of x_* for which the constant value (here, 0) is maintained is known in certain economic models as a *turnpike*.

(c) In the case $T \leqslant 2\sqrt{2}$ (short horizon), show that x_* is given by:

$$x_*(t) = \begin{cases} -t^2/2 + [(8+T^2)/(4T)]\,t - 1 & \text{if } 0 \leqslant t \leqslant T/2 \\ (t-T)^2/2 + [(8+T^2)/(4T)](t-T) + 1 & \text{if } T/2 \leqslant t \leqslant T. \end{cases}$$

We observe that there is no turnpike component in this case.

(d) Consider now the problem

$$\min \int_0^T \{\,|x'(t)|^2/2 - |x(t)|\,\}\,dt \,:\, x(0) = A,\ x(T) = B,\ x \in \text{Lip}[0,T].$$

(Note the change of sign in the second term of the integrand.) Prove that the problem has at least one solution x_*, and that any solution lies in $C^1[0,T]$.

(e) In contrast with the previous case, show that, regardless of the value of T, a solution x_* is *never* constant on a subinterval. □

18.4 Generalized Tonelli-Morrey conditions

We study in this section the problem of Bolza that consists of minimizing the functional

$$J(x) = \ell\big(x(a), x(b)\big) + \int_a^b \Lambda\big(t, x(t), x'(t)\big)\, dt$$

subject to the boundary conditions $(x(a), x(b)) \in E$. The Lagrangian $\Lambda(t,x,v)$ is assumed to be finite-valued, and LB measurable in t and (x,v); however, in contrast to the previous sections, Λ will not be taken to be locally Lipschitz. We continue to assume, however, that the endpoint cost function ℓ is locally Lipschitz, and that $E \subset \mathbb{R}^n \times \mathbb{R}^n$ is closed.

We shall define a new structural hypothesis for such Lagrangians leading to an extension of the classical necessary conditions. In contrast to Theorem 18.1, which postulated Lipschitz regularity of Λ relative to the given local minimum (this is a "solution specific" hypothesis), the new hypothesis is framed so as to *not* make specific reference to the minimizing arc.

In the following definition and subsequent theorem, the notation $\partial_P \Lambda$ and $\partial_L \Lambda$ refers to subdifferentials of the function $\Lambda(t,x,v)$ taken with respect to just the (x,v) variables.

18.11 Hypothesis. *The Lagrangian Λ satisfies the generalized Tonelli-Morrey condition: for every bounded subset S of $\mathbb{R}^n$, there exist a constant c and a summable function d such that, for almost every t, for every $(x,v) \in S \times \mathbb{R}^n$, one has*

$$\frac{|\zeta|}{1+|\psi|} \leqslant c\big(|v| + |\Lambda(t,x,v)|\big) + d(t) \;\; \forall \big(\zeta, \psi\big) \in \partial_P \Lambda(t,x,v).$$

18.12 Exercise. Let $\Lambda(t,x,v) = f(t,x) + g(t,v)$ where f is locally Lipschitz. Show that Hypothesis 18.11 holds. □

The reader will recall that an arc x is said to be admissible for the problem if it satisfies the endpoint constraints, and if the integral in $J(x)$ is well defined and finite. Let the admissible arc x_* be a strong local minimizer for the problem.

18.13 Theorem. *Under the hypotheses above, there exists an arc p which satisfies the* **Euler inclusion:**

$$p'(t) \in \operatorname{co}\big\{\omega : \big(\omega, p(t)\big) \in \partial_L \Lambda\big(t, x_*(t), x_*'(t)\big)\big\}, \;\; t \in [a,b] \text{ a.e.}$$

together with the **Weierstrass condition:** *for almost every t,*

$$\Lambda\big(t, x_*(t), v\big) - \Lambda\big(t, x_*(t), x_*'(t)\big) \geqslant \langle p(t), v - x_*'(t)\rangle \;\; \forall v \in \mathbb{R}^n$$

and the **transversality condition:**

$$\big(p(a), -p(b)\big) \in \partial_L \ell\big(x_*(a), x_*(b)\big) + N_E^L\big(x_*(a), x_*(b)\big).$$

If Λ is autonomous, then in addition the **Erdmann condition** *holds: there is a constant h such that*

$$\langle\, p(t), x_*'(t)\rangle - \Lambda\big(x_*(t), x_*'(t)\big) = h, \ \ t \in [a,b] \text{ a.e.}$$

Remark. The generalized Tonelli-Morrey structural hypothesis used here extends the classical one expressed by ($*$) in Theorem 16.13, even for smooth Lagrangians. To see this, suppose for example that Λ satisfies

$$\big|\Lambda_x(t,x,v)\big| \leqslant c\,(\,|v| + |\Lambda(t,x,v)|\,) + d(t)\big(1 + |\Lambda_v(t,x,v)|\big) \ \ \forall\,(t,x,v). \qquad (1)$$

Bearing in mind that any point (ζ,ψ) belonging to $\partial_P \Lambda(t,x,v)$ is of the form $(\Lambda_x(t,x,v), \Lambda_v(t,x,v))$, it is easy to see that the generalized Tonelli-Morrey condition 18.11 then holds. Note, however, that the structural hypothesis (1) is *weaker* than the classical condition ($*$) in Theorem 16.13: the term $|\Lambda_v|$ has migrated to the *right* side of the inequality, making the hypothesis less restrictive.

We mention, too, that the Erdmann condition is asserted here without the convexity in v that was required in Cor. 16.19. The theorem follows from later results on differential inclusions, and its proof is postponed to §25.3.

18.14 Exercise. Show that the following Lagrangian ($n = 1$) satisfies Hypothesis 18.11, but not the condition ($*$) of Theorem 16.13:

$$\Lambda(t,x,v) = \exp\big\{\big(1 + x^2 + t^2\big)v^2\big\}. \qquad \square$$

Regularity consequences. There is a close link between necessary conditions and the regularity of the solution. If we know the solution to be regular, then we can usually assert the necessary conditions; conversely, if we can write the necessary conditions, then we may be able to deduce regularity from them. The following is an illustration of this principle.

18.15 Corollary. *Under the hypotheses of* Theorem 18.13, *suppose in addition that Λ has Nagumo growth along x_*, and is bounded above on bounded subsets. Then x_* is Lipschitz.*

Proof. Theorem 18.13 provides an arc p exists which satisfies the Weierstrass condition. Let M be an upper bound on the values

$$\Lambda\big(t, x_*(t), x_*'(t)/\big(1 + |x_*'(t)|\big)\big), \ \ t \in [a,b].$$

Then, taking $v = x_*'(t)/(1 + |x_*'(t)|)$ in the Weierstrass inequality leads to

$$\Lambda\big(t, x_*(t), x_*'(t)\big) \leqslant M + |\,p(t)|\,|x_*'(t)| \text{ a.e.}$$

By the Nagumo growth hypothesis (see Def. 16.15), we then have

$$\theta\big(|x_*'(t)|\big) \leqslant M + |p(t)||x_*'(t)| \text{ a.e.}$$

Since $|p(t)|$ is bounded, and since θ has superlinear growth, this implies that x_*' is essentially bounded. □

18.16 Example. We study a problem that features a discontinuous Lagrangian and a free endpoint. It consists of minimizing

$$\int_0^\tau \big\{ |x(t)| + g\big(|x'(t)|\big) \big\}\, dt$$

over the arcs $x : [0,\tau] \to \mathbb{R}$ satisfying the endpoint constraint $x(\tau) = \beta$, with $x(0)$ being free, where g is the (discontinuous) function

$$g(r) = \begin{cases} 1 + r^2/2 & \text{if } r \neq 0 \\ 0 & \text{if } r = 0. \end{cases}$$

(Such functions arise in applications in which zero velocity has no fuel cost, whereas any nonzero velocity requires running the engine, which forces a minimal (positive) fuel consumption.) Note that the direct method is inapplicable here, since the Lagrangian $\Lambda(x,v) = |x| + g(|v|)$ is not convex in v, so that existence theory does not apply. Let us assume for now that a solution x_* exists, and let us proceed to extract information from the necessary conditions.

We observe that Λ is LB measurable, and lower semicontinuous in (x,v), and satisfies Hypothesis 18.11 (see Exer. 18.12). Thus, Theorem 18.13 is applicable, and provides a costate p satisfying (almost everywhere)

$$p(0) = 0, \ |p'(t)| \leqslant 1, \tag{2}$$

$$x_*'(t) \neq 0 \implies p(t) = x_*'(t), \quad x_*(t) \neq 0 \implies p'(t) = x_*(t)/|x_*(t)|.$$

We also deduce, for almost every t,

$$x_*'(t) \neq 0 \implies 0 - 1 - |x_*'(t)|^2/2 \geqslant -p(t)x_*'(t)$$
$$x_*'(t) = 0 \implies 1 + |v|^2/2 - 0 \geqslant p(t)v \ \ \forall v \in \mathbb{R}\backslash\{0\}$$

as a consequence of the Weierstrass condition (specialized to $v = 0$ in the first case). The first of these inequalities forces $|p(t)| \geqslant \sqrt{2}$, and the second forces the opposite, so we may restate as follows: for almost every t

$$x_*'(t) \neq 0 \implies |p(t)| \geqslant \sqrt{2} \text{ and } x_*'(t) = p(t); \ \ x_*'(t) = 0 \implies |p(t)| \leqslant \sqrt{2}. \tag{3}$$

By (2) we have $|p(t)| < \sqrt{2}$ for $t < \sqrt{2}$, whence $x_*'(t) = 0$ a.e. for $t \leqslant \sqrt{2}$. It follows that when the horizon τ satisfies $\tau \leqslant \sqrt{2}$, then x_* is identically β.

Let us now consider longer horizons $\tau > \sqrt{2}$, beginning with the case $\beta > 0$.

Then x_* must satisfy $x_*'(t) = 0$ a.e. on $[0, \sqrt{2}\,]$. Thus, there is a constant α such that

$$x_*(t) = \alpha,\ t \in [0, \sqrt{2}\,].$$

Let us consider first the case $\alpha > 0$. Then $p'(t) = 1$ for $t \in [0, \sqrt{2}\,]$, so that $p(\sqrt{2}\,) = \sqrt{2}$. After this, there is at least a small interval during which $p(t) > \sqrt{2}$ (since $p' = x_*/|x_*| = 1$ for t near $\sqrt{2}$), and in which we have $x_*'(t) \neq 0$, by (3). In that interval, we have

$$x_*'' = 1,\ x_*' = p > 0,\ p' = 1.$$

It follows that x_*, x_*' and p are strictly increasing, so that, in fact, the situation in the interval persists thereafter. We proceed to solve $x'' = 1$ with the conditions

$$x\big(\sqrt{2}\big) = \alpha,\ x'\big(\sqrt{2}\big) = p\big(\sqrt{2}\big) = \sqrt{2}$$

to reveal

$$x_*(t) = \big(t - \sqrt{2}\,\big)^2/2 + \sqrt{2}\,\big(t - \sqrt{2}\big) + \alpha,\ t \in \big[\sqrt{2}, \tau\big],$$

where $\alpha < \beta$ necessarily. Then $x_*(\tau) = \beta$ determines α; we find

$$\alpha = \beta - \big(\tau - \sqrt{2}\,\big)^2/2 - \sqrt{2}\,\big(\tau - \sqrt{2}\,\big).$$

We had imposed $\alpha > 0$, however, which (in view of the expression for α) forces $\tau < (2\beta + 2)^{1/2}$. If this fails, that is, if $\tau \geqslant (2\beta + 2)^{1/2}$, we must, once more, modify the proposed solution; we do so by introducing the possibility $\alpha = 0$. In this case, during the initial period in which $x_*(t) = 0$, the Euler inclusion now gives $|\,p'\,| \leqslant 1$, so the inequality $|\,p(t)| < \sqrt{2}$ (and $x_*(t) = 0$) can persist beyond $t = \sqrt{2}$, to some value $\sigma > \sqrt{2}$, say. Then x_* will become positive, whence $p(\sigma) = x_*'(\sigma) = \sqrt{2}$. For $t > \sigma$, we have $x_*'' = 1$; together with $x(\sigma) = 0$ and $x(\tau) = \beta$, this leads to

$$x_*(t) = (t - \sigma)^2/2 + \sqrt{2}\,(t - \sigma),$$

where $\sigma = \tau + \sqrt{2} - (2\beta + 2)^{1/2}$.

When $\beta < 0$, symmetry implies that the solution is obtained by merely changing the sign of x_* (and replacing β by $|\,\beta\,|$ in the formulas above); the solution is evidently $x_* \equiv 0$ when $\beta = 0$.

Summing up, we have used the necessary conditions to identify a unique possible solution to the problem for every value of (τ, β), without, for the moment, being able to assert that these are really solutions. The deductive method, the classical sufficiency theory, convexity arguments: all fail to apply. We complete the analysis later (see Example 19.3), when another inductive method will have become available, one that we turn to now. □

Chapter 19
Hamilton-Jacobi methods

In the preceding chapters, the predominant issues have been those connected with the *deductive method*: existence on the one hand, and the twin issues of regularity and necessary conditions on the other. We proceed now to describe the method of *verification functions*, a technique which unifies all the main inductive methods. The reader will see that this leads to a complex of ideas centered around the Hamilton-Jacobi inequality (or equation).

19.1 Verification functions

Let us illustrate the basic idea with a simple example, that of minimizing

$$J(x) = \int_0^1 \big|x'(t)\big|^2 dt$$

over the continuously differentiable functions $x : [0,1] \to \mathbb{R}^n$ satisfying the constraints $x(0) = 0$, $x(1) = \theta$, where $\theta \in \mathbb{R}^n$ is a given unit vector. Solving this problem by the deductive method requires the following steps:

- A solution $x_* \in \mathrm{AC}[0,1]$ to the problem is known to exist by Tonelli's theorem.
- A regularity theorem implies that $x_* \in \mathrm{Lip}[0,1]$ (Cor. 16.16 or Theorem 16.18).
- Since x_* is Lipschitz, the integral Euler equation applies; we deduce that x_* is C^2 and satisfies $x_*'' = 0$.
- It follows that the linear arc $x_*(t) = \theta t$ is the unique solution.

This approach involves advanced infrastructure and a lot of machinery. What if we wished to prove to a first-year undergraduate (or an economist) that x_* is indeed the solution? Here is an argument that recommends itself by its elementary nature.

F. Clarke, *Functional Analysis, Calculus of Variations and Optimal Control*, Graduate Texts in Mathematics 264, DOI 10.1007/978-1-4471-4820-3_19,

First, observe the general inequality (recall that θ is a unit vector)

$$|v|^2 = \big(|v|-1\big)^2 + 2|v| - 1 \geqslant 2|v| - 1 \geqslant 2\langle v, \theta \rangle - 1.$$

Next, proceed to replace v by $x'(t)$, where x is *any* function admissible for the problem. Now integrate over $[0,1]$ to get a lower bound on $J(x)$, for any such x:

$$J(x) = \int_0^1 \big|x'(t)\big|^2 dt \geqslant \int_0^1 \big\{2\langle x'(t), \theta \rangle - 1\big\} dt$$
$$= 2\langle x(1) - x(0), \theta \rangle - 1 = 2|\theta|^2 - 1 = 1.$$

Finally, observe that when x is x_*, the inequalities above hold with equality (prior to and after integrating); thus $J(x_*) = 1$. It follows from this completely elementary argument that x_* minimizes J subject to the given boundary conditions.[1]

We now describe in general terms the *method of verification functions* that is involved here. Recall the basic problem (P):

$$\text{minimize} \quad J(x) = \int_a^b \Lambda\big(t, x(t), x'(t)\big)\, dt$$

over the arcs $x \in \mathrm{AC}[a,b]$ satisfying $x(a) = A$, $x(b) = B$. Suppose that an admissible suspect x_* is at hand, and that our goal is to confirm that x_* solves (P).

The basic idea is very simple. Suppose that a C^1 function $\varphi(t,x)$ exists such that, for all $t \in [a,b]$ and $(x,v) \in \mathbb{R}^n \times \mathbb{R}^n$, we have

$$\Lambda(t,x,v) \geqslant \varphi_t(t,x) + \langle \varphi_x(t,x), v \rangle \tag{1}$$

Equality holds almost everywhere in (1) along x_*. (2)

This last phrase means that (1) holds with equality, for almost all t, when (t,x,v) is replaced by $(t, x_*(t), x_*'(t))$. We claim that the existence of such a φ, which we call a *verification function*, confirms the optimality of x_* for (P).

To see this, let x be any arc admissible for (P), and write the inequality (1) with $(t,x,v) = (t, x(t), x'(t))$. Then integrate over $[a,b]$ to obtain

$$J(x) = \int_a^b \Lambda\big(t, x(t), x'(t)\big)\, dt \geqslant \int_a^b \frac{d}{dt}\, \varphi\big(t, x(t)\big)\, dt = \varphi(b,B) - \varphi(a,A).$$

This last constant is therefore a lower bound on $J(x)$. But observe that if $x = x_*$, then this lower bound is attained; hence x_* solves (P). In the example above, where $\Lambda(t,x,v) = |v|^2$, the argument corresponds to taking $\varphi(t,x) = 2\langle x, \theta \rangle - t$.

[1] We could also invoke convexity in this example, depending on the student. As we shall see, the argument by convexity is a special case of the technique we are describing.

The principal questions concerning this inductive method come quickly to mind. Does there always exist a verification function φ to confirm that x_* is optimal (when it is)? And how do we find verification functions?

The following *value function* $V(\tau,\beta)$ provides a great deal of insight into both these issues:

$$V(\tau,\beta) \;=\; \inf \int_a^\tau \Lambda(t,x,x')\,dt \;:\; x\in \mathrm{AC}[a,\tau],\; x(a)=A,\; x(\tau)=\beta.$$

We refer to the problem defined by the right side as $\mathrm{P}(\tau,\beta)$. Note that (τ,β) is simply the parameter specifying the horizon and endpoint constraint of a family of problems, in which the original problem (P) = $\mathrm{P}(b,B)$ has been imbedded.

Suppose now that x_* solves (P) and that φ is a verification function for x_*. Without loss of generality, we may take $\varphi(a,A)=0$ (a verification function is indifferent to an additive constant). Let x be any arc on $[a,\tau]$ with the prescribed boundary values $x(a)=A$, $x(\tau)=\beta$. Integrate the inequality (1) evaluated along x to obtain

$$\int_a^\tau \Lambda(t,x,x')\,dt \;\geqslant\; \varphi(\tau,\beta).$$

Taking the infimum over all arcs x feasible for $\mathrm{P}(\tau,\beta)$ gives

$$V(\tau,\beta) \;\geqslant\; \varphi(\tau,\beta). \tag{3}$$

Now, for each $\tau\in(a,b]$, it follows from the optimality of x_* for (P) that x_* solves $\mathrm{P}(\tau,x_*(\tau))$. This inherited optimality for certain subproblems is sometimes called *the principle of optimality*. Interpreted in terms of V, this fact reads as follows:

$$V\big(\tau,x_*(\tau)\big) \;=\; \int_a^\tau \Lambda(t,x_*,x_*')\,dt \;=\; \varphi\big(\tau,x_*(\tau)\big),\quad a<\tau\leqslant b. \tag{4}$$

We may summarize (3) and (4) by saying that V majorizes[2] φ, while agreeing with φ along x_*. We have seen earlier that value functions are not always differentiable. The possible relevance of this fact to the verification technique is the following: if V is not differentiable, if V has a "concave corner" at a point $(\tau,\beta)=(\tau,x_*(\tau))$ (that is, a corner that has the nature of $y\mapsto -|y|$ at $y=0$), then no smooth function φ can agree with V at that point while it is majorized by V everywhere. Thus the existence of any such φ would be ruled out.[3]

On the other hand, if V happens to be smooth, our calculation suggests that φ could possibly be taken to be V itself. Of course, the very definition of V is somewhat

---

[2] Borrowing from French (or is it Latin?), we say that f *majorizes* g if $f\geqslant g$.

[3] A concave corner is precisely the type of nondifferentiability that a value function is likely to develop: consider the one-dimensional example $V(x)=\min\{xu : u\in[-1,1]\}=-|x|$.

difficult at $t = a$. It would seem natural to define $V(a,A)$ to be 0, but $V(a,\beta)$ for $\beta \neq A$ should probably be assigned the value $+\infty$.

The concern expressed above turns out to be justified. There are very regular problems (P) (with smooth Lagrangians) that generate value functions V having the type of corner alluded to. Thus the solutions of these problems do not admit (smooth) verification functions.

As for our second thought, the possibility that V itself could be a verification function, let us now confirm that there are some grounds for optimism.

19.1 Proposition. *Let Λ be continuous, and suppose that the infimum defining $V(\tau,\beta)$ is attained for each $(\tau,\beta) \in \Omega = (a,b)\times\mathbb{R}^n$, and that V is differentiable in Ω. Then $\varphi := V$ satisfies (1) for $(t,x,v) \in \Omega\times\mathbb{R}^n$, and if x_* solves (P), then V satisfies (2) as well.*

Proof. The second assertion is an immediate consequence of differentiating (4). As for the first assertion, fix (τ,β), let x solve $\mathrm{P}(\tau,\beta)$, and let any v be given. By considering the arc z that agrees with $x(t)$ for $t \leqslant \tau$ and that equals $x(\tau)+(t-\tau)v$ for $t > \tau$, we derive, for any $\delta > 0$ sufficiently small, by definition of V:

$$\begin{aligned} V\big(\tau+\delta,\beta+\delta v\big) &\leqslant \int_a^{\tau+\delta} \Lambda\big(t,z,z'\big)\,dt \\ &= \int_a^{\tau} \Lambda\big(t,x,x'\big)\,dt + \int_\tau^{\tau+\delta} \Lambda\big(t,x(\tau)+(t-\tau)v,v\big)\,dt. \end{aligned}$$

The first term on the right coincides with $V(\tau,\beta)$. If we subtract it from both sides, divide by δ, and let δ tend to zero, we obtain precisely (1), at the point (τ,β,v), and for $\varphi = V$. □

Ideally, the fact that V may be a verification function might be used as follows. We formulate a conjecture regarding the solutions of $\mathrm{P}(\tau,\beta)$ for all (τ,β); then we calculate V provisionally through substitution of the conjectured solutions. If, then, (1) and (2) hold for $\varphi = V$, the conjecture is verified. When this procedure is successful, we have actually solved a whole family of problems rather than just (P) itself.

Let us illustrate this by an example, the one with which we began this section. The parametrized problem $\mathrm{P}(\tau,\beta)$ is given by

$$\text{minimize } \int_0^\tau \big|x'(t)\big|^2 dt \text{ subject to } x(0) = 0,\; x(\tau) = \beta.$$

We suspect the solution to be the linear arc $x(t) = t\beta/\tau$, for which the integral in question turns out to be $\varphi(\tau,\beta) = |\beta|^2/\tau$. (Note that we had proposed a different verification function before.) We verify that (1) holds for $t > 0$:

$$\Lambda(t,x,v) - \varphi_t(t,x) - \langle \varphi_x(t,x), v \rangle = |v|^2 + |x|^2/t^2 - 2\langle x,v\rangle/t$$
$$= |v - x/t|^2 \geqslant 0.$$

As for (2), it holds automatically, from the way in which φ is constructed. This almost completes the proof that $x_*(t) \equiv t\,\theta$ solves P$(1,\theta)$, except for the bothersome detail of dealing with the awkward behavior of $|\beta|^2/\tau$ at $\tau = 0$.

If the argument proving the sufficiency of (1) and (2) is modified so that we integrate only over $[\varepsilon, 1]$ for small $\varepsilon > 0$, it yields

$$\int_\varepsilon^1 |x'(t)|^2 dt \geqslant 1 - \frac{|x(\varepsilon)|^2}{\varepsilon}$$

for any admissible arc x. In the limit as $\varepsilon \downarrow 0$, provided that $|x(\varepsilon)|^2/\varepsilon$ tends to zero, this inequality gives the required conclusion. If we have chosen to restrict the problem to smooth functions x, then this is indeed the case. Otherwise, we could argue that we know the solutions of P(τ,β) to be Lipschitz by our earlier regularity results; once again, this guarantees that $|x(\varepsilon)|^2/\varepsilon$ tends to zero. (The reader will note how regularity remains a pervasive issue even in this context.)

Alternatively, we could have avoided the difficulty with V at $\tau = 0$ in the above example by "backing up" the problem to the larger interval $[-\delta, 1]$ for some $\delta > 0$, as we now demonstrate. Consider the parametrized problem

$$\text{minimize } \int_{-\delta}^{\tau} |x'(t)|^2 dt \text{ subject to } x(-\delta) = -\delta\,\theta,\ x(\tau) = \beta.$$

(Notice that this extension is designed to be compatible with $x_*(t) = t\,\theta$, our putative solution.) Again we suspect that the solution is the sole admissible linear arc. Calculating the value function $V_\delta(\tau,\beta)$ on the basis of this conjecture, we obtain

$$V_\delta(\tau,\beta) = \frac{|\beta + \delta\,\theta|^2}{\tau + \delta}.$$

This function satisfies (1) and (2) for the new extended interval (for $x_*(t) = t\,\theta$ extended to $[-\delta, 1]$), and all the more for the original one. Further, observe that V_δ presents no difficulties at $\tau = 0$: there is no bothersome detail at 0 when this verification function is used.

This backing-up idea lies at the heart of a technique in the classical calculus of variations due to Hilbert and Weierstrass. A given extremal x_* on $[0,T]$ (say) is extended to one on $[-\delta, T]$. A *pencil of extremals*, all passing through $(-\delta, x_*(-\delta))$ is shown to cover a neighborhood of the graph of x_* on $[0,T]$ (if no conjugate points are present); it is said that x_* is imbedded in a *field*.

Then $V(\tau,\beta)$, defined as the integral over $[-\delta, T]$ along the (unique) element of the field passing through (τ,β), is a smooth function (locally, and for $\tau \geqslant 0$), satisfying (1) and (2) (if the Weierstrass condition holds). This celebrated theory leads to the

confirmation of x_* as a *strong* local minimum. We remark that the approach of § 14.2 can be given a similar interpretation, in which Legendre's extra term leads to a verification function (for a weak local minimum).

In this light, the verification function method may be seen as a unifying one that subsumes the classical setting, while applying to other situations, especially when it is extended to allow nonsmooth verification functions (as we shall see).

The reader will have understood that there is no uniqueness of verification functions. In the example above, using value functions, an entire family of suitable ones (parametrized by δ) was found. Note also that the verification function we gave initially in discussing the example (that is, $\varphi(t,x) = 2\langle x, \theta\rangle - t$) is seemingly unrelated to these value functions. Is it possible that it was simply devised in an *ad hoc*, trial-and-error fashion? This would not be a systematic method like value functions or field theory (or the Hamilton-Jacobi equation to be discussed presently), but it remains a persistently useful way to find verification functions.

Let us consider now sufficiency results based upon convexity, to see how they relate to verification functions. We expect that an arc x_* which solves (P) will admit a costate arc p satisfying the integral Euler equation

$$\big(p'(t), p(t)\big) = \nabla_{x,v} \Lambda\big(t, x_*(t), x_*'(t)\big) \text{ a.e.}$$

Now suppose that $\Lambda(t,\cdot,\cdot)$ is convex for each t. Then the subdifferential inequality for convex functions asserts that we have, for almost every t:

$$\Lambda(t,x,v) - \Lambda\big(t, x_*(t), x_*'(t)\big) \geqslant \\ \langle p'(t), x - x_*(t)\rangle + \langle p(t), v - x_*'(t)\rangle \;\; \forall\,(x,v) \in \mathbb{R}^n \times \mathbb{R}^n.$$

If we proceed to define

$$\varphi(t,x) = \langle p(t), x - x_*(t)\rangle + \int_a^t \Lambda\big(s, x_*(s), x_*'(s)\big)\, ds,$$

then by construction, (2) holds. Furthermore, the last inequality above is precisely (1), thus demonstrating that φ is a verification function for x_*, one that was constructed with the help of the necessary conditions. We conclude, therefore, that the convex case, too, is subsumed by the verification method. (We now confess to the reader that the function $\varphi(t,x) = 2\langle x, \theta\rangle - t$ used at the very beginning of the section was found in this way.)

The Hamiltonian H corresponding to Λ is the function

$$H(t,x,p) = \sup_{v \in \mathbb{R}^n} \{\langle p, v\rangle - \Lambda(t,x,v)\}.$$

In terms of H, the fact that inequality (1) holds for all v can be expressed as a *Hamilton-Jacobi inequality*

$$\varphi_t(t,x) + H\big(t,x,\varphi_x(t,x)\big) \leqslant 0.$$

In the idealized context of Prop. 19.1, if one assumes in addition that solutions x_* are smooth, the proof of the proposition implies that V satisfies this inequality with *equality*. As we shall see in §19.3, it is this value function that yearns to be the solution, though it is regrettably prone to being nondifferentiable. In fact, the Hamilton-Jacobi equation does not generally admit a smooth global solution. Note, however, that we only need *sub*solutions (of the inequality) in particular applications where we wish to apply the verification function technique.

The following identifies a context in which *locally Lipschitz* verification functions can be used; it allows us to integrate the Hamilton-Jacobi inequality when it holds in the almost everywhere sense.

19.2 Proposition. *Let $\Lambda(t,x,v)$ be* LB measurable*, continuous in (t,x), as well as bounded below on bounded sets. Let Ω be an open subset of $(a,b)\times\mathbb{R}^n$, and W a subset of $\mathbb{R}^n$. Suppose that $\varphi:\Omega\to\mathbb{R}$ is a locally Lipschitz function satisfying the Hamilton-Jacobi inequality in the almost everywhere sense*:

$$\varphi_t(t,x) + \langle\,\varphi_x(t,x),v\,\rangle \leqslant \Lambda(t,x,v)\ \ \forall\, v\in W,\ \ (t,x)\in\Omega \text{ a.e.} \tag{5}$$

Then, for any $x\in Lip[a,b]$ satisfying

$$\big(t,x(t)\big)\in\Omega,\ \ x'(t)\in W,\ \ t\in(a,b) \text{ a.e.},$$

we have

$$J(x) \geqslant \limsup_{\varepsilon\downarrow 0}\,\big\{\,\varphi\big(b-\varepsilon,x(b-\varepsilon)\big)-\varphi\big(a+\varepsilon,x(a+\varepsilon)\big)\big\}. \tag{6}$$

We omit the proof, as a more general result will be proved later (see Prop. 24.5).

Of course, the intended use of Prop. 19.2 is to extend the verification procedure to nonsmooth functions φ, as follows. Consider the minimization of $J(x)$ over the Lipschitz arcs x satisfying certain boundary conditions, and suppose that for all such arcs x, the upper limit appearing in (6) is no less than a certain value J_0. Then J_0 is a lower bound on the value of the problem. If we exhibit an admissible arc x_* for which $J(x_*) = J_0$, then it follows that x_* is a solution of the problem.

Because of its adaptability, we view the verification function approach as a technique more than a theory. Accordingly, we proceed in the rest of the section (and in the next) to illustrate its use in varied contexts.

19.3 Example. We continue the analysis of the problem studied in Example 18.16, which had been left incomplete. In order to implement the verification function method, it is useful to know the cost $\varphi(t,x)$ that corresponds to the arcs that appear in our conjecture, where (t,x) corresponds to the parameter values (τ,β) in the endpoint constraint of the problem. Substitution and routine calculation yield

$$\varphi(t,x) = \begin{cases} t|x| & \text{if} \quad 0 \leqslant t \leqslant \sqrt{2} \\ t(|x|+1) - t^3/6 - \sqrt{2}^{\,3}/3 & \text{if} \quad \sqrt{2} \leqslant t \leqslant \sqrt{2|x|+2} \\ \sqrt{2|x|+2}^{\,3}/3 - \sqrt{2}^{\,3}/3 & \text{if} \quad t \geqslant \sqrt{2|x|+2}\,. \end{cases}$$

It is not hard to see that this function is locally Lipschitz on $\mathbb{R}_+ \times \mathbb{R}$ (with the help of Exer. 13.18, say). Let us now verify the Hamilton-Jacobi inequality (5). Consider first the case $0 < t < \sqrt{2}$ and $x \neq 0$. If $v = 0$, we have

$$\Lambda(x,v) - \varphi_t(t,x) - \langle \varphi_x(t,x), v \rangle = |x| - |x| = 0.$$

If $v \neq 0$, then (using $t < \sqrt{2}$)

$$\begin{aligned} \Lambda(x,v) - \varphi_t(t,x) - \langle \varphi_x(t,x), v \rangle &= |x| + 1 + v^2/2 - |x| - t\,x\,v/|x| \\ &\geqslant 1 + v^2/2 - \sqrt{2}\,|v| = \big(|v|/\sqrt{2} - 1\big)^2 \geqslant 0. \end{aligned}$$

Consider next the case $\sqrt{2} < t < \sqrt{2|x|+2}$ and $x \neq 0$. If $v = 0$, we have

$$\Lambda(x,v) - \varphi_t(t,x) - \langle \varphi_x(t,x), v \rangle = |x| - (|x|+1) + t^2/2 = -1 + t^2/2 \geqslant 0,$$

since $t > \sqrt{2}$. If $v \neq 0$, then

$$\begin{aligned} \Lambda(x,v) - \varphi_t(t,x) - \langle \varphi_x(t,x), v \rangle &= |x| + 1 + v^2/2 - (|x|+1) + t^2/2 - t\,x\,v/|x| \\ &\geqslant v^2/2 + t^2/2 - t|v| = \big(|v| - t\big)^2/2 \geqslant 0. \end{aligned}$$

The remaining case ($t > \sqrt{2|x|+2}$) is treated similarly.

Since φ is continuous, and because we have $\varphi(0,\cdot) = 0$, Prop. 19.2 yields

$$J(x) \geqslant \varphi(\tau,\beta) - \varphi(0,x(0)) = \varphi(\tau,\beta)$$

for any Lipschitz arc x joining the points $(0,x(0))$ and (τ,β). But this lower bound is achieved by our proposed solutions: they were used to calculate φ. Thus, their optimality is confirmed.

Note that the argument in this example required the boundary condition $\varphi(0,\cdot) = 0$; this corresponds to $x(0)$ being free in the problem. Note, too, that the analysis has led us to solve a *family* of problems; this is a characteristic of the method. □

The next case in point illustrates the use of verification functions in the presence of auxiliary constraints (as does the next section).

19.4 Example. (A problem with a differential constraint)
On a given interval $[0,T]$, it is required to build up the quantity $x(t)$ of stock of a given material good from its initial value $x(0) = 0$ to its required final level Q. The corresponding problem is the following:

$$\min \int_0^T \left\{ x'(t)^2 + 4\,x(t) \right\} dt \ : \ x(0) = 0,\ x(T) = Q,$$

where the first term in the integral is a production cost, and the second reflects the cost of storing the stock. Note the conflicting factors: we would prefer to postpone production as much as possible (to reduce storage costs), but rapid production near the end of the period would be very costly (because of the quadratic term).

The nature of the model imposes an auxiliary constraint: $x'(t) \geqslant 0$ (since we do not envisage negative production rates). If we ignore this differential constraint and proceed to solve this (convex) problem, the Euler equation $x'' = 2$ leads to the unique solution

$$x(t) = t^2 + \big(Q/T - T\big)\,t.$$

This function has a nonnegative derivative on $[0,T]$ if $T \leqslant \sqrt{Q}$; it follows that it solves the problem in that case. If $T > \sqrt{Q}$, however, x' becomes negative in the interval; it is therefore unacceptable as a solution. We must find the solution which respects the differential constraint.

For longer horizons, we may guess that there will be an initial period $[0,\sigma]$ during which $x = 0$, following which $x(t)$ will be strictly positive. On $[\sigma,T]$ the Euler equation applies, so that $x(t)$ has the form $(t-\sigma)^2 + c\,(t-\sigma)$ for some constant c.

The Erdmann condition $x'^2 - 4\,x = h$ holds on $[0,T]$ (even when the differential constraint is present; see Cor. 16.19), and we must have $h = 0$ (in light of the initial period). We deduce from this $c = 0$, and then $x(T) = Q$ gives $\sigma = T - \sqrt{Q}$. In summary, here is our educated guess:

$$x(t) = \begin{cases} 0 & \text{if} \quad 0 \leqslant t \leqslant T - \sqrt{Q} \\ (t-\sigma)^2 & \text{if} \quad T - \sqrt{Q} \leqslant t \leqslant T. \end{cases}$$

How do we confirm this conjecture? We proceed to calculate by substitution the cost $\varphi(T,Q)$ associated to the proposed strategy:

$$\varphi(T,Q) = \begin{cases} 2\,QT - T^3/3 + Q^2/T & \text{if} \quad T \leqslant \sqrt{Q} \\ (8/3)\,Q^{3/2} & \text{if} \quad T > \sqrt{Q}\,. \end{cases}$$

Then φ is locally Lipschitz on $(0,\infty)\times\mathbb{R}_+$ (by Exer. 13.18). Now we proceed to check (5) for $t > 0$ and $v \geqslant 0$. The restriction to $x \geqslant 0$ may also be made because admissible arcs are nonnegative, so that the values of a verification function when $x < 0$ are irrelevant. When $0 < t < \sqrt{x}$, we find

$$\begin{aligned} \Lambda(x,v) - \varphi_t(t,x) - \langle\, \varphi_x(t,x), v\,\rangle &= v^2 - 2\,v\,(t + x/t) + t^2 + 2\,x + x^2/t^2 \\ &= \big(v + (t + x/t)\big)^2 \geqslant 0. \end{aligned}$$

When $t > \sqrt{x}$, we find

$$\Lambda(x,v) - \varphi_t(t,x) - \langle \varphi_x(t,x), v \rangle = \big(v + 2\sqrt{x}\big)^2 \geqslant 0.$$

Now let x be any Lipschitz admissible arc for the problem (thus, $x \geqslant 0$). Integration of the inequality (5), with the help of Prop. 19.2, leads to

$$J(x) \geqslant \varphi(T,Q) - \liminf_{\varepsilon \downarrow 0} \varphi\big(\varepsilon, x(\varepsilon)\big).$$

Since x is Lipschitz, there is a constant K such that $|x(\varepsilon)| \leqslant K\varepsilon$. This (together with the definition of φ) implies that the lower limit in the inequality above is zero. Thus, the lower bound $\varphi(T,Q)$ is obtained for $J(x)$, one that is attained for the proposed solution, which is therefore confirmed as optimal.

Note that the formal proof is elementary and requires no theory; it suffices to be armed with the function φ (which was, admittedly, identified with the help of the necessary conditions). We remark that for this problem (in contrast to that of the preceding example), a deductive approach is also available (see Exer. 21.19). □

19.2 The logarithmic Sobolev inequality

We give in this section an elementary proof of a celebrated inequality. The analysis will illustrate the use of the verification function technique for isoperimetric problems, and in proving inequalities. The *logarithmic Sobolev inequality*, published by Leonard Gross in 1976, states that for any continuously differentiable complex-valued function u on $\mathbb{R}^n$, we have

$$\int_{\mathbb{R}^n} |u(x)|^2 \ln |u(x)|\, d\mu_n(x) \leqslant \frac{1}{2} \|\nabla u\|^2 + \|u\|^2 \ln \|u\|, \tag{1}$$

where μ_n denotes Gaussian measure:

$$d\mu_n(x) = \pi^{-n/2} \exp\big(-|x|^2\big)\, dx,$$

and $\|\cdot\|$ denotes the norm on the Hilbert space $L^2(\mathbb{R}^n, d\mu_n)$:

$$\|v\|^2 = \int_{\mathbb{R}^n} |v(x)|^2\, d\mu_n(x).$$

This inequality, which has important applications in quantum field theory and elsewhere, is usually proved via lengthy probabilistic arguments. We now give a proof using only elementary calculus.

We begin by making several reductions. First, it can be shown that we need prove (1) only for the case $n = 1$; one verifies that the general case then follows (as noted by Gross) by induction on n. Next, observe that (1) is homogeneous with respect to scaling in u, so we may assume hereafter that $\|u\| = 1$. We may also assume that

$\| u' \| < \infty$, or else there is nothing to prove. Under these assumptions, the change of variable $u(t) = v(t)\exp(t^2/2)$ leads to the equivalent formulation:

$$\int_{\mathbb{R}} |v(t)|^2\,dt = \sqrt{\pi} \implies \int_{\mathbb{R}} \Big(\frac{1}{2}|v'(t)|^2 - |v(t)|^2 \ln|v(t)| \Big)\,dt \geqslant \frac{\sqrt{\pi}}{2}. \quad (2)$$

Since $||v|'| \leqslant |v'|$ a.e., we may assume in proving (2) that v is nonnegative and real-valued, as well as locally Lipschitz. For technical reasons, we shall assume that $v(t) > 0\ \forall t \in \mathbb{R}$; this is justified by a simple density argument. Finally, it is convenient to split (2) into two half-line problems, each equivalent to

$$\int_0^\infty v(t)^2 dt = \frac{\sqrt{\pi}}{2} \Longrightarrow \int_0^\infty \Big(\frac{1}{2}v'(t)^2 - v(t)^2 \ln v(t) \Big)\,dt \geqslant \frac{\sqrt{\pi}}{4}. \quad (3)$$

We summarize to this point: it is a matter of proving (3) when v is a locally Lipschitz function satisfying

$$v(t) > 0, \ \int_0^\infty v(t)^2 dt = \frac{\sqrt{\pi}}{2}, \ \int_0^\infty v'(t)^2 dt < \infty. \quad (4)$$

For $s, r > 0$, we define a function V as follows:

$$V(s,r) = \big\{ g s^2 + r\big(1 - g^2 - 2\ln s\big)\big\}/2,$$

where $g(s,r) = h^{-1}(r/s^2)$, and where h itself is given by

$$h(t) = e^{t^2} \int_t^\infty e^{-\tau^2} d\tau.$$

It is not difficult to check that h is strictly decreasing on $\mathbb{R}$: this is evident for $t \leqslant 0$, and for $t > 0$, the inequality $h'(t) < 0$ is equivalent to

$$\int_t^\infty e^{-\tau^2} d\tau < e^{-t^2}/(2t) = \int_t^\infty (-e^{-\tau^2}/(2\tau))' d\tau,$$

which in turn follows from the inequality $e^{-\tau^2} < (-e^{-\tau^2}/(2\tau))'$. It also follows that the strictly decreasing function $h^{-1} : (0,\infty) \to \mathbb{R}$ satisfies

$$h^{-1}(\sqrt{\pi}/2) = 0, \ (h^{-1})'(\rho) = \{2\rho\, h^{-1}(\rho) - 1\}^{-1},$$

from which the partial derivatives of V may then be computed:

$$V_s = g s, \quad V_r = -g^2/2 - \ln s.$$

Let us set $F(s,u) = u^2/2 - s^2 \ln s$. Then for $s > 0$ and $u \in \mathbb{R}$, one has

$$V_s u - V_r s^2 + F(s,u) = \frac{1}{2}(u + gs)^2 \geqslant 0. \quad (5)$$

Accordingly, if v satisfies (4), we have, using (5),

$$\frac{d}{dt}V\Big(v(t),\int_t^\infty v(\tau)^2d\tau\Big) = V_s\,v'(t)-V_r\,v(t)^2 \geqslant -F\big(v(t),v'(t)\big)$$

and

$$\begin{aligned}\int_0^\infty F\big(v(t),v'(t)\big)\,dt \;&\geqslant -\int_0^\infty \frac{d}{dt}V\Big(v(t),\int_t^\infty v(\tau)^2d\tau\Big)\,dt\\ &\geqslant V\Big(v(0),\frac{\sqrt{\pi}}{2}\Big) - \liminf_{t\to\infty} V\Big(v(t),\int_t^\infty v(\tau)^2d\tau\Big).\end{aligned}$$

Now we have

$$V_s(s,\sqrt{\pi}/2) = h^{-1}\big(\sqrt{\pi}/(2s^2)\big)s,$$

which is negative for $0<s<1$, and positive for $s>1$. This implies that the function $s\mapsto V(s,\sqrt{\pi}/2)$ attains a minimum over $(0,\infty)$ at $s=1$, the minimum being $V(1,\sqrt{\pi}/2)=\sqrt{\pi}/4$. It follows that the desired inequality (3) will have been established once we have proved the following lemma.

Lemma. *If v satisfies* (4) *then*

$$\liminf_{t\to\infty}\;V\Big(v(t),\int_t^\infty v(\tau)^2d\tau\Big) \leqslant 0.$$

Proof. It is readily checked that $h(\tau)<1/\tau$ for all $\tau>0$. Hence $h^{-1}(t)<1/t$ for all $t>0$, and $g(s,r)\,s^2<s^4/r$. Similarly,

$$h(\tau)<\sqrt{\pi}\exp(\tau^2)\;\;(\tau\leqslant 0)\implies h^{-1}(t)\leqslant -\sqrt{\ln\big(t/\sqrt{\pi}\big)}\;\;(t\geqslant\sqrt{\pi}),$$

and hence

$$g(s,r)^2 \geqslant \ln r-\ln s^2-\ln\sqrt{\pi}\;\text{ for }\;\sqrt{\pi}\,s^2\leqslant r.$$

Evidently $g(s,r)^2\geqslant 0$ if $r<\sqrt{\pi}\,s^2$, and so

$$r\,(1-g^2-2\ln s)\leqslant \max\big\{\,r\big(1+\ln\sqrt{\pi}-\ln r\big),\;\sqrt{\pi}\,s^2\big(1-\ln s^2\big)\,\big\}\;\;\forall r,s>0.$$

We deduce from these estimates

$$V(s,r)\leqslant \frac{s^4}{2r}+\frac{1}{2}\max\big\{\,r\big(1+\ln\sqrt{\pi}-\ln r\big),\;\sqrt{\pi}\,s^2\big(1-\ln s^2\big)\,\big\}. \tag{6}$$

If v satisfies (4), then the three functions

$$s=v(t),\;\; r=\int_t^\infty v(\tau)^2d\tau,\;\text{and}\;\;\varepsilon=\int_t^\infty v'(\tau)^2d\tau$$

all tend to zero as t tends to infinity. Moreover,

$$s^4 = v(t)^4 \leqslant \left(2\int_t^\infty v(\tau)\big|v'(\tau)\big|\,d\tau\right)^2 \leqslant 4r\varepsilon$$

by Hölder's inequality. Thus, by (6), $\liminf_{t\to\infty} V(s,r) \leqslant 0$, as required. □

Remark. The reader undoubtedly suspects (correctly) that the eminently useful function V that lies at the heart of the proof is a value function. If we consider the parametrized problem

$$\text{minimize } J(v) = \int_0^\infty F(v,v')\,dt \;\text{ subject to } \int_0^\infty v^2 dt = r \text{ and } v(0) = s,$$

then the multiplier rule for the isoperimetric problem (Theorem 14.21) suggests that a minimum is attained at the function $v(t,s,r) = c\exp(-(t+b)^2)$, where c and b are determined by the two constraints involving s and r. We cannot be sure about this, since we are unable (initially) to affirm that a solution exists, and since the multiplier rule was not proved for infinite horizon problems such as the current one.

However, all's fair in provisionally calculating a verification function: the end will justify the means (if we are right). Thus, we proceed to define $V(s,r) = J(v(\cdot,s,r))$ (precisely the function V of the proof), and the verification of the inequality (5) simply shows that V satisfies the appropriate Hamilton-Jacobi inequality. [4]

19.3 The Hamilton-Jacobi equation

We study in this section the following Cauchy problem for the Hamilton-Jacobi equation:

$$\textbf{(HJ)}\quad \begin{cases} u_t + H(x,u_x) = 0,\ (t,x) \in \Omega := (0,\infty)\times\mathbb{R}^n \\ u(0,x) = \ell(x),\ x\in\mathbb{R}^n. \end{cases}$$

The *Hamiltonian* $H : \mathbb{R}^n\times\mathbb{R}^n \to \mathbb{R}$ is given, as well as the function $\ell : \mathbb{R}^n \to \mathbb{R}$ defining the boundary values at time $t = 0$.

Classical solutions. The function u is a *classical solution* of (HJ) if u belongs to the class $C^1(\overline{\Omega})$ and satisfies[5]

$$u_t(t,x) + H\big(x,u_x(t,x)\big) = 0,\ (t,x)\in\Omega \;\text{ and }\; u(0,x) = \ell(x)\ \forall x\in\mathbb{R}^n.$$

---

[4] The results of this section appear in [1].

[5] The function u belongs to $C^1(\overline{\Omega})$ if it is continuously differentiable in Ω, and if u, as well as all its first-order partial derivatives, admit continuous extensions to $\overline{\Omega}$.

Of course, this pointwise solution concept is natural and desirable. However, nonlinear differential equations (even physically meaningful ones) do not generally admit smooth, globally defined solutions (as Legendre belatedly discovered).

The following Lagrangian Λ, obtained from H via the Legendre-Fenchel transform, plays a central role in clarifying the issue:

$$\Lambda(x,v) \;=\; \max\{\langle v,p\rangle - H(x,p) \,:\, p \in \mathbb{R}^n\}. \tag{1}$$

We shall be supposing throughout that the Hamiltonian H is continuous, and convex with respect to p. Then, as we know from the theory of conjugate convex functions (see Theorem 4.21), applying the transform to Λ brings us back to H:

$$H(x,p) \;=\; \max\{\langle p,v\rangle - \Lambda(x,v) \,:\, v \in \mathbb{R}^n\}. \tag{2}$$

In a setting such as this, it was known to our ancestors (informally, at least) that the natural solution of (HJ) is the function

$$u_*(\tau,\beta) \;=\; \min\; \ell\big(x(0)\big) + \int_0^\tau \Lambda\big(x(t),x'(t)\big)\,dt\,,$$

where the minimum is taken over the arcs x on $[0,\tau]$ satisfying $x(\tau)=\beta$ (the initial value $x(0)$ being free). As a value function, however, u_* is generally lacking in smoothness; in that case, it cannot be a classical solution.

19.5 Exercise. Let u be a classical solution of (HJ) relative to the restricted domain $D = (0,T)\times\mathbb{R}^n$, where H and Λ are continuous functions satisfying (2). Prove that $u \leqslant u_*$ on D. □

19.6 Example. Consider (HJ) with the data $n=1$ and

$$H(x,p) \;=\; p^2 e^x/4\,,\quad \ell(x) \;=\; -e^{-x}.$$

We calculate $\Lambda(x,v) = e^{-x}v^2$. In an attempt to identify the solution to the problem above that defines $u_*(\tau,\beta)$, we write the Euler equation $x'' = (x')^2/2$. Combined with the transversality condition (see Theorem 18.1), this implies

$$p(0) \;=\; e^{-x(0)} \;=\; 2e^{-x(0)}x'(0),$$

which yields $x'(0) = 1/2$. Together with the boundary condition $x(\tau)=\beta$, we are led to propose the solution

$$x(t) = -2\ln(ct+k),\;\text{ where } c = e^{-\beta/2}/(\tau-4),\;\; k=-4c.$$

The resulting cost from this arc is readily calculated:

$$u(\tau,\beta) \;=\; 4e^{-\beta}/(\tau-4).$$

We do not necessarily affirm that this is really u_*, but certainly we have $u_* \leqslant u$, since u is the cost of certain admissible arcs, and u_* is the least cost.

Since $u(\tau,\beta) \downarrow -\infty$ as $\tau \uparrow 4$, it follows that u_*, and (by Exer. 19.5) any classical solution defined at least on $(0,4)\times\mathbb{R}$, cannot be appropriately defined beyond $t=4$. The moral that we wish to draw is that certain growth conditions must be imposed on the data if we are to have globally defined solutions of (HJ) (smooth or not). □

Almost everywhere solutions. Due to the possible nonexistence of classical solutions, we are motivated to seek a generalization that will authorize nonsmooth functions as solutions. The reader has met such a concept in connection with verification functions (see Prop. 19.2), as well as in Example 11.24, one that depends on the fact that a locally Lipschitz function is differentiable almost everywhere. We define u to be an *almost everywhere solution* if it is locally Lipschitz on $\overline{\Omega}$, satisfies the boundary condition $u(0,\cdot) = \ell(\cdot)$, and satisfies the Hamilton-Jacobi equation at almost all points where u is differentiable:

$$u_t(t,x)+H\big(x,u_x(t,x)\big) = 0, \quad (t,x)\in\Omega \text{ a.e.}$$

The difficulty with the almost everywhere approach (from the point of view of the elegance of the theory) is that such solutions are not unique in general; there will be many solutions of that type.[6]

19.7 Example. Consider (HJ) with data $H(x,p) = |p|^2-1$, $\ell(x)\equiv 0$. There exists a classical solution which is bounded below: the function $u(t,x) = t$. (Results to come will tell us it is unique.) However, there also exist different solutions in the almost everywhere sense, for example $u(t,x) = \min(t,|x|)$. That this (Lipschitz) function defines such a solution is easy to see, since, almost everywhere, x is nonzero and $|x| \neq t$, and then u coincides locally with either t, x, or $-x$; in each case, u satisfies locally $u_t+|u_x|^2-1 = 0$. □

Proximal solutions. The moral of the example above is that uniqueness, if it is to be asserted, will require that we look at the points of nondifferentiability of the candidate solutions, and not simply ignore them as being of measure zero. A natural way to do this is to involve a subdifferential in the definition of the extended solution concept.

We say that a function $u : [0,\infty)\times\mathbb{R}^n \to \mathbb{R}$ is a *proximal solution* to (HJ) provided

- u is locally Lipschitz on $[0,\infty)\times\mathbb{R}^n$ and satisfies $u(0,x) = \ell(x) \;\; \forall x\in\mathbb{R}^n$;
- $(t,x)\in(0,\infty)\times\mathbb{R}^n,\ (\theta,\zeta)\in\partial_P u(t,x) \implies \theta+H(x,\zeta) = 0.$

[6] In the context of verification functions, however (where, furthermore, we are dealing with an inequality rather than an equality), non uniqueness is rather desirable, since the more verification functions there are, the easier (presumably) it will be to find one.

19.8 Exercise.

(a) Suppose that u is a classical solution of (HJ). Prove that u is a proximal solution of (HJ).

(b) Conversely, let u be a proximal solution of (HJ), where u lies in $C^1(\overline{\Omega})$ and H is continuous. Prove that u is a classical solution of (HJ).

(c) Show that a proximal solution of (HJ) is also an almost everywhere solution.

(d) Show that the almost everywhere solution $u(t,x) = \min(t,|x|)$ adduced in Example 19.7 *fails* to be a proximal solution of (HJ). □

19.9 Exercise. Let u be a proximal solution to (HJ) for the boundary function ℓ_1, and let v be a proximal solution for the boundary function ℓ_2. Prove that $\min(u,v)$ is a proximal solution to (HJ) for the boundary function $\ell := \min(\ell_1, \ell_2)$. □

An existence and uniqueness theorem. The following hypotheses are made concerning the data of the problem.[7]

19.10 Hypothesis.

(a) *ℓ is locally Lipschitz and bounded below*;

(b) *$H(x,p)$ is locally Lipschitz, and is convex as a function of p for each x*;

(c) *H has superlinear growth in p in the following sense*:

$$\lim_{|p|\to\infty} H(x,p)/|p| = \infty \text{ uniformly for } x \text{ in bounded sets};$$

(d) *There exist positive constants C, $\kappa > 1$ and $\sigma < \kappa$ such that*

$$H(x,p) \leqslant C|p|^{\kappa}\big(1+|x|\big)^{\sigma} \;\; \forall (x,p) \in \mathbb{R}^n \times \mathbb{R}^n.$$

19.11 Theorem. *Under* Hypothesis 19.10, *there is a unique proximal solution u_* of* (HJ) *which is bounded below.*

The proof will make use of the Lagrangian Λ defined by (1).

19.12 Exercise. Prove that Λ is locally Lipschitz, and coercive in the following sense: for certain positive constants $c_1, r > 1$, and $s < r$ we have

$$\Lambda(x,v) \geqslant \frac{c_1|v|^r}{(1+|x|)^s} \;\; \forall x, v \in \mathbb{R}^n.$$ □

[7] There are other sets of hypotheses that would serve here; these are intended to be indicative.

A characterization of the solution. In proving Theorem 19.11, we shall also derive the following characterization of u_*:

For each $(\tau,\beta) \in [0,\infty) \times \mathbb{R}^n$*,* $u_*(\tau,\beta)$ *is the minimum in the following problem* $\mathrm{P}(\tau,\beta)$ *in the calculus of variations* :

$$\mathrm{P}(\tau,\beta) \quad \begin{cases} \text{minimize } J_\tau(x) = \ell\big(x(0)\big) + \displaystyle\int_0^\tau \Lambda\big(x(t), x'(t)\big)\,dt \\ \text{over the arcs } x \text{ on } [0,\tau] \text{ satisfying } x(\tau) = \beta. \end{cases}$$

The hypotheses imply that this problem admits a solution (another exercise in the direct method; see Exer. 16.11). The necessary conditions of Theorem 18.1 can be helpful in solving $\mathrm{P}(\tau,\beta)$, which features a free initial value and a (generally) nonsmooth Lagrangian. We remark that under the hypotheses of Theorem 19.11, even when the data ℓ and H are smooth, the solution u_* of (HJ) may not be smooth (see Exer. 21.33).

The proof of Theorem 19.11 is given in the next section.

19.13 Exercise. Show that Theorem 19.11 applies to the following problem, and find the corresponding solution:

$$\textbf{(HJ)} \quad \begin{cases} u_t + |u_x|^2 = 0, \ (t,x) \in (0,\infty) \times \mathbb{R}^n \\ u(0,x) = |x|^2, \ x \in \mathbb{R}^n. \end{cases}$$ □

19.14 Exercise. Find the unique nonnegative (proximal or classical) solution to the boundary-value problem

$$\textbf{(HJ)} \quad \begin{cases} u_t + |u_x|^2 - 1 = 0, \ (t,x) \in (0,\infty) \times \mathbb{R}^n \\ u(0,x) = |x|, \ x \in \mathbb{R}^n. \end{cases}$$ □

19.15 Exercise. (The Hopf-Lax formula) In the context of Theorem 19.11, suppose in addition that H is differentiable and independent of x. Let u_* be the solution of (HJ) provided by the theorem. Prove that we have

$$u_*(\tau,\beta) = \min_{\alpha \in \mathbb{R}^n} \big\{ \ell(\alpha) + \tau\Lambda\big((\beta-\alpha)/\tau\big) \big\}, \ \tau > 0, \beta \in \mathbb{R}^n.$$ □

Viscosity solutions. We now forge a link between the proximal solution studied above, and the well-known *viscosity solutions* of (HJ). This refers to a continuous function u on $[0,\infty) \times \mathbb{R}^n$ satisfying the boundary condition, such that, for every $(t,x) \in (0,\infty) \times \mathbb{R}^n$, we have

$$\theta + H(x,\zeta) \geqslant 0 \ \forall (\theta,\zeta) \in \partial_D u(t,x), \ \ \theta + H(x,\zeta) \leqslant 0 \ \forall (\theta,\zeta) \in \partial^D u(t,x).$$

Here, $\partial_D u$ refers to the Dini (or viscosity) subdifferential (see §11.4), and $\partial^D u$, the Dini superdifferential, is defined by $\partial^D u = -\partial_D(-u)$. The connection to proximal solutions will be made by means of the next result.

19.16 Proposition. *Let $g : \mathbb{R}^n \to \mathbb{R}$ be convex and $f : \Omega \to \mathbb{R}$ locally Lipschitz, where Ω is an open subset of $\mathbb{R}^n$. Then*

$$g(\zeta) \leqslant 0 \;\; \forall\, \zeta \in \partial^D f(x) \;\; (x \in \Omega) \iff g(\zeta) \leqslant 0 \;\; \forall\, \zeta \in \partial_P f(x) \;\; (x \in \Omega).$$

Proof. It turns out to be convenient to prove somewhat more than what is stated, namely the equivalence of the following six properties:

$$\begin{array}{ll}
(1)\;\; g(\zeta) \leqslant 0 \;\; \forall\, \zeta \in \partial^D f(x),\, x \in \Omega & (2)\;\; g(\zeta) \leqslant 0 \;\; \forall\, \zeta \in -\partial_L(-f)(x),\, x \in \Omega \\
(3)\;\; g(\zeta) \leqslant 0 \;\; \forall\, \zeta \in \partial_C f(x),\, x \in \Omega & (4)\;\; g(\zeta) \leqslant 0 \;\; \forall\, \zeta \in \partial_L f(x),\, x \in \Omega \\
(5)\;\; g(\zeta) \leqslant 0 \;\; \forall\, \zeta \in \partial_D f(x),\, x \in \Omega & (6)\;\; g(\zeta) \leqslant 0 \;\; \forall\, \zeta \in \partial_P f(x),\, x \in \Omega.
\end{array}$$

That (1) $\Longrightarrow$ (2) follows because $\partial_L(-f)$ is obtained from $\partial_D(-f) = -\partial^D f$ via a sequential closure operation (see Cor. 11.47), and since g (being convex and finite) is continuous. We see that, in fact, (1) and (2) are equivalent.

Lemma. *Let $x \in \Omega$. Then*

$$\max_{\partial_L f(x)} g(\zeta) \;=\; \max_{\partial_C f(x)} g(\zeta) \;=\; \max_{-\partial_L(-f)(x)} g(\zeta).$$

Proof. The first equality holds because $\partial_C f = \operatorname{co} \partial_L f$ and g is convex; the maximum of a convex function over a compact set and over its convex hull coincide. The second is essentially the same fact as the first, since $\partial_C f = -\partial_C(-f)$.

Returning to the proof of the proposition, we observe that, by the lemma, (2) $\Longrightarrow$ (3). That (3) $\Longrightarrow$ (4) $\Longrightarrow$ (5) $\Longrightarrow$ (6) is evident, since the sets are becoming smaller. Now suppose that (6) holds. Since $\partial_D f$ can be approximated by $\partial_P f$ (see Theorem 11.45), we deduce (5). Because $\partial_L f$ is obtained from $\partial_D f$ by sequential closure (Cor. 11.47), we then deduce (4). By the lemma, this yields in turn (2), or equivalently, (1). □

19.17 Corollary. *Let $H(x,p)$ be continuous in (x,p) and convex in p. A locally Lipschitz function u on $(0,\infty)\times\mathbb{R}^n$ satisfies*

$$\theta + H(x,\zeta) = 0 \;\; \forall\, (\theta,\zeta) \in \partial_P u(t,x),\;\; (t,x) \in (0,\infty)\times\mathbb{R}^n \tag{3}$$

if and only if it satisfies

$$\theta + H(x,\zeta) \geqslant 0 \;\; \forall\, (\theta,\zeta) \in \partial_D u(t,x),\;\; (t,x) \in (0,\infty)\times\mathbb{R}^n \tag{4}$$

$$\text{and}\;\; \theta + H(x,\zeta) \leqslant 0 \;\; \forall\, (\theta,\zeta) \in \partial^D u(t,x),\;\; (t,x) \in (0,\infty)\times\mathbb{R}^n. \tag{5}$$

Proof. If (3) holds, then (4) holds as well, since $\partial_D u$ is approximated by $\partial_P u$, in the sense of Theorem 11.45. In addition, (5) follows from the proposition. Conversely, if (4) and (5) hold, then, for any $(t,x) \in (0,\infty)\times\mathbb{R}^n$, for any $(\theta,\zeta) \in \partial_P u(t,x)$, we have

$$\theta + H(x,\zeta) \geqslant 0,$$

by (4), since $\partial_P u \subset \partial_D u$. But we also deduce $\theta + H(x,\zeta) \leqslant 0$, by applying the proposition to (5). Thus $\theta + H(x,\zeta) = 0$. □

The corollary implies that, in the context of Theorem 19.11, the function u_* is the unique locally Lipschitz viscosity solution of (HJ) that is bounded below. The topic of generalized solutions of the Hamilton-Jacobi equation is revisited in a different context in Exer. 26.30.

19.4 Proof of Theorem 19.11

We note without proof a useful bound which results from the coercivity of Λ established in Exer. 19.12.

19.18 Proposition. *For each bounded subset S of $[0,\infty)\times\mathbb{R}^n$ and number N, there exists M such that*

$$(\tau,\beta) \in S,\ x \text{ an arc on } [0,\tau],\ x(\tau) = \beta,\ J_\tau(x) \leqslant N \implies \int_0^\tau |x'(t)|^r dt \leqslant M.$$

19.19 Corollary. *There exists Q such that, for any $(\tau,\beta) \in S$, for any solution $x_{\tau,\beta}$ of the problem* $\mathrm{P}(\tau,\beta)$*, we have $|x_{\tau,\beta}(t)| \leqslant Q$ for all $t \in [0,\tau]$.*

Proof. This follows from the easily-proved fact that $u_*(\tau,\beta)$, as the value of the problem $\mathrm{P}(\tau,\beta)$, is bounded above on bounded sets. □

We shall require the following estimate:

19.20 Proposition. *If $S \subset [0,\infty)\times\mathbb{R}^n$ is bounded, there exists m such that*

$$(\tau,\beta) \in S \implies \text{ess min}\,\{\,|x'_{\tau,\beta}(t)| : t \in [0,\tau]\,\} \leqslant m.$$

Proof. We have seen above that $|x_{\tau,\beta}(t)|$ is uniformly bounded for $(\tau,\beta) \in S$ and $t \in [0,\tau]$, by a constant Q. In light of Exer. 19.12, we have $\Lambda(x,v) \geqslant c|v|^r$ for some constant $c > 0$, whenever $|x| \leqslant Q$. Let K_ℓ be a Lipschitz constant for ℓ on the appropriate bounded set. If y denotes the arc which is identically β, the inequality $J_\tau(x_{\tau,\beta}) \leqslant J_\tau(y)$ (which reflects the optimality of $x_{\tau,\beta}$) leads to (putting $x = x_{\tau,\beta}$)

$$\int_0^\tau c|x'|^r\,dt \leqslant \ell(\beta)-\ell\big(x(0)\big)+\tau\Lambda(\beta,0) \leqslant K_\ell\,|x(0)-x(\tau)|+\tau\Lambda(\beta,0)$$
$$\leqslant K_\ell \int_0^\tau |x'|\,dt+\tau\Lambda(\beta,0).$$

In turn this gives

$$\int_0^\tau \big\{c|x'|^r-K_\ell|x'|\big\}\,dt \leqslant \tau\Lambda(\beta,0),$$

which implies

$$\text{ess min}\,\big\{c\big|x'(t)\big|^r-K_\ell\big|x'(t)\big| : t\in[0,\tau]\big\} \leqslant \Lambda(\beta,0),$$

which yields the desired result since Λ is bounded on bounded sets. □

We know that, for some constant $\lambda=\lambda_{\tau,\beta}$, the solution $x=x_{\tau,\beta}$ satisfies

$$\Lambda\big(x(t),x'(t)\big)-x'(t)\bullet\zeta(t) = \lambda,\ t\in[0,\tau] \text{ a.e.},$$

where $\zeta(t)\in\partial_v\Lambda(x(t),x'(t))$ a.e. (see Step B of the proof of Theorem 16.18; this is an Erdmann condition). The preceding proposition now leads to:

19.21 Exercise. Prove that if $S\subset[0,\infty)\times\mathbb{R}^n$ is bounded, there exists λ_S such that

$$(\tau,\beta)\in S \implies |\lambda_{\tau,\beta}| \leqslant \lambda_S.$$

Deduce from this that there exists a constant E_S such that

$$(\tau,\beta)\in S,\ t\in[0,\tau] \implies \big|x'_{\tau,\beta}(t)\big| \leqslant E_S.$$ □

Armed as we now are with the fact that $|x'_{\tau,\beta}(t)|$ is uniformly bounded for (τ,β) in bounded sets, we can easily show:

19.22 Proposition. *For each bounded subset S of $[0,\infty)\times\mathbb{R}^n$ there exists K_S such that*

$$(\tau,\alpha),\,(\tau,\beta)\in S \implies \big|u_*(\tau,\alpha)-u_*(\tau,\beta)\big| \leqslant K_S\,|\alpha-\beta|.$$

Proof. We know that $|x_{\tau,\beta}(t)|$ and $|x'_{\tau,\beta}(t)|$ are uniformly bounded for (τ,β) in S and $t\in[0,\tau]$. Let K_ℓ and K_Λ be Lipschitz constants for ℓ and Λ on a suitable neighborhood of the bounded set in question. If we define

$$y(t) = x_{\tau,\beta}(t)+(\tau-t)(\alpha-\beta),\ t\in[0,\tau],$$

then (setting $x=x_{\tau,\beta}$)

$$
\begin{aligned}
u_*(\tau,\alpha) \leqslant J_\tau(y) \ \ &\text{(by the definition of } u_* \text{ as a value)}\\
&= \ell\big(x(0)+\tau(\alpha-\beta)\big)+\int_0^\tau \Lambda\big(x+(\tau-t)(\alpha-\beta),x'-(\alpha-\beta)\big)\,dt\\
&\leqslant \ell\big(x(0)\big)+K_\ell|\tau||\alpha-\beta|+\int_0^\tau\{\Lambda(x,x')+K_\Lambda(\tau+1)|\alpha-\beta|\}\,dt\\
&\leqslant \ell\big(x(0)\big)+\int_0^\tau \Lambda(x,x')\,dt+\big\{K_\ell\,|\,\tau\,|+K_\Lambda\,\tau(\tau+1)\big\}|\,\alpha-\beta\,|\\
&\leqslant u_*(\tau,\beta)+K_S|\,\alpha-\beta\,|
\end{aligned}
$$

for a suitable choice of K_S. □

The reader is asked to complete the proof that u_* is locally Lipschitz:

19.23 Exercise. Prove that if $S \subset [0,\infty)\times\mathbb{R}^n$ is bounded, then the function u_* is also Lipschitz on S relative to the τ variable. Deduce that u_* is locally Lipschitz. □

Next, we verify the proximal Hamilton-Jacobi equation.

19.24 Proposition. *Let* $(\theta,\zeta)\in\partial_P u_*(\tau,\beta)$, *where* $\tau>0$. *Then* $\theta+H(\beta,\zeta)=0$.

Proof. There exists $\sigma\geqslant 0$ such that, for all (t,x) in a neighborhood of (τ,β), the proximal subgradient inequality holds:

$$u_*(t,x)-u_*(\tau,\beta)\;\geqslant\;\theta(t-\tau)+\zeta\bullet(x-\beta)-\sigma\big\{\,|t-\tau|^2+|x-\beta\,|^2\big\}.$$

Fix any $v\in\mathbb{R}^n$, and let y be the arc that extends $x_{\tau,\beta}$ beyond τ with constant derivative v. Then, for all $\varepsilon>0$ sufficiently small,

$$
\begin{aligned}
J_{\tau+\varepsilon}(y)-J_\tau\big(x_{\tau,\beta}\big)\;&=\int_\tau^{\tau+\varepsilon}\Lambda\big(y(t),v\big)\,dt\;\geqslant u_*\big(\tau+\varepsilon,\,\beta+\varepsilon v\big)-u_*(\tau,\beta)\\
&\geqslant\;\theta\,\varepsilon+\varepsilon\,\zeta\bullet v-\sigma\,\varepsilon^2\{\,1+|\,v\,|^2\,\}.
\end{aligned}
$$

Dividing by ε and letting ε decrease to 0 reveals $\theta+\zeta\bullet v-\Lambda(\beta,v)\leqslant 0$. Since H is the conjugate of Λ and v is arbitrary, we derive

$$H(\beta,\zeta)=\sup_v\;\zeta\bullet v-\Lambda(\beta,v)\;\leqslant\;-\theta\,.$$

It suffices now to exhibit one value $v=v_0$ for which $\zeta\bullet v_0-\Lambda(\beta,v_0)\geqslant-\theta$, which we proceed to do. We have (setting $x=x_{\tau,\beta}$), for all ε sufficiently small,

$$
\begin{aligned}
-\int_{\tau-\varepsilon}^{\tau}\Lambda(x,x')\,dt\;&=\int_0^{\tau-\varepsilon}\Lambda(x,x')\,dt-\int_0^\tau\Lambda(x,x')\,dt\\
&=u_*\big(\tau-\varepsilon,x(\tau-\varepsilon)\big)-u_*(\tau,\beta)\\
&\geqslant\;-\varepsilon\,\theta+\zeta\bullet\big(x(\tau-\varepsilon)-x(\tau)\big)-\sigma\big\{\,\varepsilon^2+|x(\tau-\varepsilon)-x(\tau)|^2\,\big\}.
\end{aligned}
$$

Dividing across by ε, we obtain

$$\theta + \frac{1}{\varepsilon}\int_{\tau-\varepsilon}^{\tau}\big\{\,\zeta \bullet x' - \Lambda(x,x')\,\big\}\,dt \geqslant -\sigma\big\{\,\varepsilon + |x(\tau-\varepsilon)-x(\tau)|^2/\varepsilon\,\big\}.$$

Note that the right side tends to 0 as $\varepsilon \to 0$, since x is Lipschitz. The integral on the left, for given ε, is bounded above by a term of the form

$$\zeta \bullet v_\varepsilon - \Lambda\big(x(t_\varepsilon), v_\varepsilon\big),$$

for some t_ε in $[\,\tau-\varepsilon,\tau\,]$ and v_ε in kB, where k is a Lipschitz constant for x. (This follows from the mean value theorem 10.17 for the generalized gradient, along with the gradient formula 10.27.) Selecting a subsequence $\varepsilon_i \downarrow 0$ for which v_{ε_i} converges to a limit v_0, we deduce $\theta + \zeta \bullet v_0 - \Lambda(\beta, v_0) \geqslant 0$. □

Since ℓ is bounded below and Λ is nonnegative, it follows from its very definition as a value function that u_* is bounded below. There remains to show that u_* is the *unique* proximal solution to (HJ) that is bounded below. We begin with:

19.25 Proposition. *Let u be a proximal solution to* (HJ). *Then $u \leqslant u_*$.*

Proof. Fix any $\tau > 0$ and $\beta \in \mathbb{R}^n$. We prove that $u(\tau,\beta) \leqslant u_*(\tau,\beta)$. To this end, let $x = x_{\tau,\beta}$ and consider the (continuous) function

$$f(t) = u\big(t, x(t)\big) - \ell\big(x(0)\big) - \int_0^t \Lambda\big(x(s), x'(s)\big)\,ds.$$

Note that $f(0) = 0$ and $f(\tau) = u(\tau,\beta) - u_*(\tau,\beta)$. It suffices therefore to prove that $f(\tau) \leqslant 0$. But this is a direct consequence of Prop. 19.2, since u is an almost everywhere solution of (HJ) (see Exer. 19.8). □

The final step in the proof of Theorem 19.11 is the following:

19.26 Proposition. *Let u be a proximal solution to* (HJ) *that is bounded below. Then $u \geqslant u_*$.*

Proof. Fix $(\tau,\beta) \in (0,\infty)\times\mathbb{R}^n$. We shall prove that $u_*(\tau,\beta) \leqslant u(\tau,\beta)$.

By hypothesis, there exists a constant $\mu > 0$ such that $u \geqslant -\mu$.

Lemma 1. *There exists L such that, for any x on $[\,0,\tau\,]$ with $x(\tau) = \beta$, for any $\delta \in (0,\tau\,]$,*

$$\int_{\tau-\delta}^{\tau} \Lambda(x,x')\,dt \;\leqslant\; u(\tau,\beta)+\mu \implies |x(t)-\beta| \;\leqslant\; L\,|t-\tau| \;\; t \in [\,\tau-\delta,\tau\,].$$

The lemma, whose relevance will become apparent, follows from elementary arguments (using Hölder's inequality and the coercivity of Λ) that lead to a uniform estimate of the type

$$\int_t^\tau |x'(s)|\,ds \leqslant C.$$

Lemma 2. *There exists an arc x on the interval $[0,\tau]$ satisfying $x(\tau)=\beta$ and*

$$u\big(t,x(t)\big) + \int_t^\tau \Lambda\big(x(s),x'(s)\big)\,ds \leqslant u(\tau,\beta)\ \ \forall t \in [0,\tau]. \tag{1}$$

Proof. We define the compact "triangle" Δ by

$$\Delta = \big\{ (t,x) : t \in [0,\tau],\, |x-\beta| \leqslant L|t-\tau| \big\}$$

(where L is given by Lemma 1), as well as a bigger triangle Δ^+ containing Δ:

$$\Delta^+ = \big\{ (t,x) : t \in [0,\tau],\, |x-\beta| \leqslant L|t-\tau-1| \big\}.$$

Let K be a Lipschitz constant for u on the set Δ^+, and let R be an upper bound for $|x|$ over all points $(t,x) \in \Delta^+$. Due to the coercivity of Λ, there is a constant M sufficiently large so that the following implication holds:

$$\begin{aligned} &|x| \leqslant R,\ |p| \leqslant K \Longrightarrow \\ &\qquad \max_{|v| \leqslant M} \big\{ \langle p,v\rangle - \Lambda(x,v) \big\} = \max_{v \in \mathbb{R}^n} \big\{ \langle p,v\rangle - \Lambda(x,v) \big\} = H(x,p). \end{aligned} \tag{2}$$

Now let N satisfy

$$|x| \leqslant R,\ |v| \leqslant M \Longrightarrow |\Lambda(x,v)| \leqslant N.$$

The data above will help us define a system for which Theorem 12.11 can be invoked, yielding the lemma. This involves treating t as a component of the state, introducing an additional state coordinate y to absorb Λ into the dynamics, and reversing time. (So the notation is a wee bit complex.) We define

$$F(t,x,y) = \big\{ (-1,w,r) : |w| \leqslant M,\ \Lambda(x,-w) \leqslant r \leqslant N \big\}.$$

Then it can be verified that F satisfies Hypothesis 12.1 for $\widetilde{\Delta} := \Delta^+ \times \mathbb{R}$, as well as linear growth. We further define

$$\Omega_+ = \big(\operatorname{int} \Delta^+\big) \times \mathbb{R},\ \ \varphi(t,x,y) = u(t,x) + y.$$

We claim that the system (φ,F) is weakly decreasing on Ω_+, a claim that will be established by verifying the proximal criterion of Theorem 12.11. (This is where the proximal inequality $\theta + H(x,\zeta) \geqslant 0$ satisfied by u will play a role; only the opposite inequality has actually been needed so far.)

Accordingly, let (θ,ζ,γ) belong to $\partial_P \varphi(t,x,y)$, where $(t,x,y) \in \Omega_+$. It follows that $\gamma = 1$ and $(\theta,\zeta) \in \partial_P u(t,x)$; thus, $\theta + H(x,\zeta) = 0$. We also have $|\zeta| \leqslant K$ (the Lipschitz constant for u), whence

$$\begin{aligned} h_F\big(t,x,y,\theta,\zeta,\gamma\big) &= \min\big\{\langle(\theta,\zeta,\gamma),(-1,w,r)\rangle : (-1,w,r)\in F(t,x,y)\big\} \\ &= \min\big\{-\theta+\langle\zeta,w\rangle+\Lambda(x,-w) : |w|\leqslant M\big\} \\ &= -\theta-\max\big\{\langle\zeta,-w\rangle-\Lambda(x,-w) : w\in\mathbb{R}^n\big\} \text{ (by (2))} \\ &= -\theta-H(x,\zeta) = 0. \end{aligned}$$

By Theorem 12.11 and the definition of weak decrease, there is a trajectory for F, originating at the point $(\tau,\beta,0)$, having the form $(\tau-t,\tilde{x}(t),y(t))$, maximally defined on an interval $[0,T)$ relative to Ω_+, such that, for t in this interval, we have

$$\begin{aligned} \varphi\big(\tau-t,\tilde{x}(t),y(t)\big) &= u\big(\tau-t,\tilde{x}(t)\big)+\int_0^t\Lambda\big(\tilde{x}(s),-\tilde{x}\,'(s)\big)\,ds \\ &\leqslant \varphi(\tau,\beta,0) = u(\tau,\beta). \end{aligned}$$

The change of variables $t = t-\tau$, together with the definition $x(t) := \tilde{x}(\tau-t)$, yields an arc satisfying the inequality in (1) on any interval $[\delta,\tau]$ $(\delta>0)$ during which $(t,x(t))$ remains in int Δ^+. However, the inequality of Lemma 1 holds, as is easily seen (in view of the lower bound on u), so that $(t,x(t))$ never leaves Δ as we decrease t from its initial value of τ; thus, $(t,x(t))$ lies in int Δ^+ $\forall\, t\in(0,\tau)$. The lemma is proved. □

It is now a simple matter to deduce Prop. 19.26. Taking $t = 0$ in Lemma 2, and bearing in mind how $u_*(\tau,\beta)$ is defined, we obtain

$$u_*(\tau,\beta) \leqslant \ell\big(x(0)\big)+\int_0^\tau\Lambda\big(x(s),x'(s)\big)\,ds \leqslant u(\tau,\beta),$$

as required. □

This completes the proof of Theorem 19.11.

Chapter 20

Multiple integrals

This chapter is an introduction to the multiple integral calculus of variations. Throughout the discussion, Ω denotes a nonempty open bounded subset of $\mathbb{R}^n$, and the problem under consideration is the minimization of the functional

$$J(u) = \int_{\Omega} F\big(x, u(x), Du(x)\big)\, dx$$

over a class X of real-valued functions u defined on Ω. The values of the admissible functions are prescribed on the boundary of Ω:

$$u(x) = \varphi(x) \;\; \forall x \in \Gamma := \partial\Omega\,,$$

where $\varphi : \Gamma \to \mathbb{R}$ is a given function.

The reader may recognize this as an extension to multiple integrals of the problem in the calculus of variations that was so very thoroughly studied in the preceding chapters. Accordingly, we refer to it as the **basic problem** in (or relative to) the class X. It turns out that, as was the case for single integrals ($n = 1$), it is perhaps the choice of X that makes the greatest difference in the ensuing theory.

We may as well admit what the reader has already observed: there has been a severe discontinuity in the notation from the single integral case. (It is best, at times, to bow to tradition.) The underlying domain $[a,b]$ has become Ω, and n now refers to its dimension. Furthermore, the Lagrangian, the independent variable, and the competing functions have mutated as follows:

$$\Lambda \longrightarrow F\,,\;\; t \longrightarrow x\,,\;\; x(t) \longrightarrow u(x).$$

We now use the generic notation x, u, and z for the three variables that appear in the Lagrangian. Thus we write $F(x,u,z)$, for $(x,u,z) \in \mathbb{R}^n \times \mathbb{R} \times \mathbb{R}^n$.

F. Clarke, *Functional Analysis, Calculus of Variations and Optimal Control*,
Graduate Texts in Mathematics 264, DOI 10.1007/978-1-4471-4820-3_20,

20.1 The classical context

We begin again with the case in which the competing functions are smooth, so smooth that we need not really think about that aspect of the problem. To be precise, we take u in the class $X = C^2(\overline{\Omega})$, where this means that u is twice continuously differentiable in Ω, and that u, as well as all its partial derivatives up to order two, admit continuous extensions to $\overline{\Omega}$. We also assume that the Lagrangian F is itself C^2.

In this agreeable setting, the cost functional is well defined everywhere on the underlying space, and the basic necessary condition asserts:

20.1 Theorem. *Any solution u_* of the basic problem in the class $C^2(\overline{\Omega})$ satisfies the Euler equation*:

$$\begin{aligned}\operatorname{div} \nabla_z F\big(x, u_*(x), Du_*(x)\big) &:= \sum_{i=1}^{n} \frac{\partial}{\partial x_i}\big\{ F_{z_i}\big(x, u_*(x), Du_*(x)\big)\big\} \\ &= F_u\big(x, u_*(x), Du_*(x)\big) \ \ \forall x \in \Omega .\end{aligned}$$

The notation 'div' refers here to the divergence operator. The proof below introduces notation such as $F(*)$, a convenient shorthand for $F(x, u_*(x), Du_*(x))$.

Proof. We merely sketch the standard variational argument. It suffices to prove that, for any $\psi \in C_c^2(\Omega)$, we have

$$\int_{\Omega} \{\operatorname{div} \nabla_z F(*) - F_u(*)\}\, \psi(x)\, dx = 0. \tag{1}$$

($C_c^2(\Omega)$ is the space of functions of class C^2 in Ω which have compact support in Ω, and which are therefore equal to zero near the boundary.) To see this, observe that the function g defined by $g(\lambda) = J(u_* + \lambda\,\psi)$ has a minimum at $\lambda = 0$, whence

$$g'(0) = \int_{\Omega} \{F_u(*)\psi(x) + \langle F_z(*), D\psi(x)\rangle\}\, dx = 0.$$

The classical divergence theorem transforms this conclusion into (1). □

Many of the single integral considerations we have seen earlier have their counterparts in the multiple integral case. For example, it is clear that a suitably defined *local minimum* would suffice to obtain the Euler equation above. Another example concerns the principle of least action, which extends to multiple integrals and the mechanics of continua, as we now illustrate.

20.2 Example. (Vibrating string) A homogeneous extensible string of mass m is stretched between two points that are distance ℓ apart, and a vibration is then induced. The profile of the string at every time instant t is described by a function

$x \mapsto u(t,x)$, $0 \leqslant x \leqslant \ell$. For fixed t, the following expressions give the kinetic energy K and the potential energy V of this physical system in terms of u:

$$K = \frac{m}{2\ell}\int_0^\ell u_t(t,x)^2\,dx, \quad V = \tau\int_0^\ell \left(\sqrt{1+u_x(t,x)^2}-1\right)dx,$$

where the (known) elasticity parameter $\tau > 0$ relates potential energy to the tension of the string, as measured by the extent that it is stretched from equilibrium.

The action between two instants t_1 and t_2 is defined as usual by

$$\int_{t_1}^{t_2}(K-V)\,dt = \int_{t_1}^{t_2}\int_0^\ell \left\{m u_t(t,x)^2/(2\ell) - \tau\sqrt{1+u_x(t,x)^2}+\tau\right\}dx\,dt.$$

The principle of least action asserts that the actual motion u of the string will minimize this functional (subject to the appropriate boundary conditions). Note that this leads to a special case of the basic problem, one in which Ω is a rectangle in $\mathbb{R}^2$.

Exercise. Show that the Euler equation for this functional is

$$u_{tt} = \frac{\ell\tau}{m}\frac{\partial}{\partial x}\left(\frac{u_x}{\sqrt{1+u_x^2}}\right).$$

For vibrations of small amplitude (that is, when $|u_x|$ is small), the integrand in the expression above for the potential energy V is approximated by $(1/2)u_x^2$. If this approximation is used in writing the action, then the resulting Euler equation is the much more tractable (since *linear*) differential equation

$$u_{tt} = \frac{\ell\tau}{m}u_{xx}.$$

This is the vibrating string equation that one finds in classical mechanics. □

20.3 Example. (The Dirichlet principle) A celebrated example in the multiple integral calculus of variations arises in connection with Laplace's equation, in dimension $n = 2$, in which we seek a solution $u = u(x,y)$ of the following partial differential equation with boundary condition:

$$u_{xx}+u_{yy} = 0 \ \text{ for } (x,y) \in \Omega, \quad u|_\Gamma = \varphi.$$

Consider the basic problem in the calculus of variations with data

$$F(u,z) = \tfrac{1}{2}|z|^2 = \tfrac{1}{2}\left(z_1^2+z_2^2\right), \ J(u) = \int_\Omega \tfrac{1}{2}\left(u_x^2+u_y^2\right)dx\,dy.$$

Then, the Euler equation of Theorem 20.1 becomes

$$\operatorname{div} Du = \frac{\partial}{\partial x}u_x + \frac{\partial}{\partial y}u_y = u_{xx}+u_{yy} = 0,$$

that is, Laplace's equation. Thus, it would appear, it suffices to find a function u minimizing $J(u)$ (subject to the boundary condition) in order to have a solution of Laplace's equation. (This is an observation due to Dirichlet.)

The issue is in fact quite complicated (and much studied), since the existence of a minimum depends on the choice of function class, the nature of Ω, and the properties of the boundary function φ. It turns out that we cannot progress in this direction if we limit attention to functions u in the class $C^2(\overline{\Omega})$. Accordingly, we shall return later to the Dirichlet problem. □

20.4 Example. (The problem of Plateau) This is the problem of finding *minimal surfaces*. It involves the minimization of the surface area functional

$$A(u) = \int_{\Omega} \sqrt{1 + |Du(x)|^2}\, dx$$

relative to the hypersurfaces $u = u(x)$ $(x \in \Omega)$ having prescribed boundary values: $u = \varphi$ on Γ. □

It turns out that minimal surfaces will generally be nonsmooth. Once again, therefore, we are motivated to consider the basic problem with functions that are less regular, a topic that we turn to now.

20.2 Lipschitz solutions

The classical context of the preceding section does not allow for nonsmooth solutions, even though these can manifest themselves physically. Further, it is not suitable for developing existence theory. An appealing class of functions offering us some relief on both these fronts is the space $\text{Lip}(\Omega)$ of functions that are globally Lipschitz on Ω.

A Lipschitz function u is differentiable almost everywhere, by Rademacher's theorem, so that the term Du in the functional J can still be interpreted as the usual derivative. Furthermore, a function u in $\text{Lip}(\Omega)$ admits natural values on the boundary of Ω. More precisely, since Ω is bounded and u is uniformly continuous, u has a unique extension to an element of $\text{Lip}(\overline{\Omega})$. We see therefore that the boundary condition $u = \varphi$ on Γ retains an unambiguous (pointwise) meaning.

Another goal of ours is to relax the smoothness hypotheses on the Lagrangian F itself. We shall do this as follows:

20.5 Hypothesis. *F is measurable in t and locally Lipschitz in (u, z), in the following sense: for every bounded subset S of $\mathbb{R} \times \mathbb{R}^n$, there exists $k \in L^1(\Omega)$ such that, for almost every $x \in \Omega$, we have*

$$\big|F(x,u_1,z_1)-F(x,u_2,z_2)\big| \leqslant k(x)\big|(u_1,z_1)-(u_2,z_2)\big|, \ (u_i,z_i) \in S \ (i=1,2).$$

In this Lipschitz setting, we obtain a generalized form of the Euler equation:

20.6 Theorem. (Clarke) *Let u_* solve the basic problem relative to Lip(Ω), under* Hypothesis 20.5. *Then there exist summable functions $p:\Omega\to\mathbb{R}^n$ and $q:\Omega\to\mathbb{R}$ such that*

$$\big(q(x),p(x)\big) \in \partial_C F\big(x,u_*(x),Du_*(x)\big) \text{ a.e. } x\in\Omega, \tag{E1}$$

and such that

$$\int_\Omega q(x)\,\psi(x)+\langle p(x),D\psi(x)\rangle\,dx = 0 \ \forall\,\psi\in Lip_0(\Omega). \tag{E2}$$

The generalized gradient $\partial_C F$ appearing above is taken with respect to the (u,z) variables; $\mathrm{Lip}_0(\Omega)$ refers to the functions in $\mathrm{Lip}(\Omega)$ which vanish on $\partial\Omega$.

Proof. Fix any $\psi\in X=\mathrm{Lip}_0(\Omega)$. For any $\lambda>0$ (sufficiently small, if we have just a local minimum), optimality implies

$$\int_\Omega \frac{F(x,\,u_*+\lambda\,\psi,\,Du_*+\lambda D\psi)-F(*)}{\lambda}\,dx \geqslant 0.$$

Taking the upper limit as $\lambda\downarrow 0$ yields (with the help of Fatou's lemma):

$$\int_\Omega F^\circ\big(*;(\psi,D\psi)\big)\,dx \geqslant 0,$$

where F° denotes the generalized directional derivative with respect to the (u,z) variables. We have, for almost every x,

$$F^\circ\big(*;(\psi(x),D\psi(x))\big) = \max_{(q,p)\in\partial_C F(*)} (q,p)\bullet\big(\psi(x),D\psi(x)\big),$$

so we deduce

$$\min_{\psi\in X}\int_\Omega \max_{(q,p)\in\partial_C F(*)} (q,p)\bullet\big(\psi(x),D\psi(x)\big)\,dx = 0.$$

Let Y denote the set of all measurable functions $x\mapsto\big(q(x),p(x)\big)$ on Ω satisfying (E1). By measurable selection theory (see Exer. 13.24), there exists, for given ψ, an element (q,p) of Y such that the maximum in the preceding integral is attained almost everywhere at x. This implies

$$\min_{\psi\in X}\max_{(q,p)\in Y}\int_\Omega \big(q(x),p(x)\big)\bullet\big(\psi(x),D\psi(x)\big)\,dx = 0.$$

Every element (q,p) of Y satisfies $|(q(x),p(x)|\leqslant k(x)$ for a certain summable function k, by Prop. 10.5. It follows that Y is weakly compact in $L^1(\Omega)$ (see Prop. 6.17).

Accordingly, by the minimax theorem 4.36, we may commute min and max in the last equation above, which gives the conclusion of the theorem. □

The weak Euler equation. Together, (E1) and (E2) constitute the Euler equation in the current setting. When q and p satisfy (E2), it is customary to say that $\operatorname{div} p = q$ in the *weak sense* (or in the *sense of distributions*). For that reason, the term *weak Euler equation* is used. In the presence of this convention, (E1) and (E2) can be expressed as an inclusion for p, one that is reminiscent of the Euler inclusion encountered in Theorem 18.1, or in Exer. 18.4:

$$\big(\operatorname{div} p(x), p(x)\big) \in \partial_C F\big(x, u_*(x), Du_*(x)\big) \text{ a.e.}$$

When F is locally Lipschitz (in all its variables), it follows from the proof of Theorem 20.6 that q and p are essentially bounded. When F is continuously differentiable, the weak Euler equation (Lipschitz version) may be written

$$\int_\Omega \{ F_u(*)\psi(x) + \langle F_z(*), D\psi(x) \rangle \} \, dx = 0 \;\; \forall \, \psi \in \operatorname{Lip}_0(\Omega).$$

If in addition the mapping $x \mapsto F_z(x, u_*(x), Du_*(x))$ is locally Lipschitz, the weak Euler equation implies the classical Euler equation of Theorem 20.1 (almost everywhere in Ω). This can be proved with the help of the following integration by parts formula in $\operatorname{Lip}(\Omega)$ (whose proof we omit).

20.7 Theorem. (Green's theorem for Lipschitz functions)
Let $f \in Lip(\Omega)$ and $g \in Lip_0(\Omega)$. Then, for each $i = 1, 2, \dots, n$, we have

$$\int_\Omega f(x) \, \frac{\partial g}{\partial x_i}(x) \, dx = -\int_\Omega \frac{\partial f}{\partial x_i}(x) \, g(x) \, dx.$$

It should not be inferred from the presence of the word "weak" above that the necessary condition itself is defective. In fact, as we now prove, the weak Euler equation is a sufficient condition for optimality when F is convex in (u, z) (compare with Theorem 15.9).

20.8 Proposition. *Let the Lagrangian $F(x, u, z)$ be measurable in t and convex with respect to (u, z), and let $u_* \in Lip(\Omega)$ be such that $J(u_*)$ is defined and finite. Suppose that there exist summable functions p and q satisfying* (E1) *and* (E2)*. Then, for any $u \in Lip(\Omega)$ having the same values as u_* on $\partial\Omega$, we have $J(u) \geqslant J(u_*)$. If the convexity of $F(x, \cdot)$ is strict for each x, and if $u \neq u_*$, then strict inequality holds.*

Proof. The hypotheses imply that F is continuous with respect to (u, z) (Cor. 2.35), so that F is LB measurable (Prop. 6.35). Fix any $\psi \in \operatorname{Lip}_0(\Omega)$. Then the function

$$x \mapsto q(x)\,\psi(x) + \langle p(x), D\psi(x) \rangle$$

is summable, and bounded above (for almost every x) by the directional derivative $F'(*;(\psi(x), D\psi(x))$ (by Cor. 4.26). But this in turn is bounded above by

$$F\big(x, u_*(x)+\psi(x), Du_*(x)+D\psi(x)\big) - F\big(x, u_*(x), Du_*(x)\big),$$

by Prop. 2.22. This reveals that $J(u_*+\psi)$ is well defined and satisfies

$$J(u_*+\psi) \geqslant J(u_*).$$

Since $\psi \in \mathrm{Lip}_0(\Omega)$ is arbitrary, we obtain the first conclusion. The second follows from the strict convexity of $J(\cdot)$. □

The Dirichlet problem: weak solutions. Consider the n-dimensional Dirichlet boundary value problem (D) that consists of finding a function u that satisfies

$$\triangle u(x) = 0 \ (x \in \Omega), \quad u|_\Gamma = \varphi, \tag{D}$$

where $\triangle$ is the *Laplacian operator*:

$$\triangle u = \sum_{i=1}^{n} \frac{\partial^2}{\partial x_i^2} u.$$

(This extends Example 20.3, in which n equals 2.) We take $\varphi : \mathbb{R}^n \to \mathbb{R}$ to be Lipschitz. In the context of functions u belonging to $\mathrm{Lip}(\Omega)$, a *weak solution* of (D) means that, besides satisfying the boundary condition, u satisfies

$$\int_\Omega Du(x) \bullet D\psi(x)\,dx = 0 \ \forall\,\psi \in \mathrm{Lip}_0(\Omega).$$

The following facts clarify the notion of weak solution.

20.9 Exercise.

(a) Show that if $u \in C^2(\overline{\Omega})$ satisfies the boundary condition and the equation $\triangle u(x) = 0$ in Ω, then u is a weak solution of the problem (D) above.

(b) Let $\hat{u} \in \varphi + \mathrm{Lip}_0(\Omega)$. Prove that $\hat{u}$ is a weak solution of (D) if and only if $\hat{u}$ minimizes the (Dirichlet) functional

$$u \mapsto \int_\Omega |Du(x)|^2 dx$$

relative to $u \in \varphi + \mathrm{Lip}_0(\Omega)$.

(c) Deduce that there is at most one weak solution of (D). □

It follows from the above that if (D) admits a classical solution in $C^2(\overline{\Omega})$, it is the unique weak solution. Furthermore, a minimizer of the Dirichlet functional provides a weak solution of the Dirichlet problem. Notice, however, that at this point we have

no existence theory allowing us to assert that the Dirichlet functional admits a minimizer, Lipschitz or otherwise. We turn now to a celebrated approach to precisely that issue.

20.3 Hilbert-Haar theory

It turns out that, by exploiting certain properties of the boundary condition, existence of solutions to the basic problem in $\mathrm{Lip}(\Omega)$ can be asserted in some situations. This is the subject of the classical Hilbert-Haar approach, initiated by Hilbert, and then contributed to by many others.

The method will now be described for the case in which the Lagrangian $F(x,u,z)$ is a convex function of z, with no dependence on x and u: this assumption is made throughout this section. (It follows that F is continuous.) This is a significant simplification of the basic problem, of course; note, however, that this class of problems does include the Dirichlet principle and the problem of Plateau.[1]

We begin by showing that affine functions automatically enjoy a minimality property in the current context.

20.10 Proposition. *Let $u_*(x) = \langle v_*, x\rangle + c$ be an affine function. Then, for every $u \in \mathrm{Lip}(\Omega)$ satisfying $u|_\Gamma = u_*|_\Gamma$, we have*

$$J(u) = \int_\Omega F\big(Du(x)\big)\,dx \geqslant J(u_*).$$

Proof. Let $\zeta \in \partial F(v_*)$ (which is nonempty by Cor. 4.7). Then, by the subdifferential inequality, we have

$$J(u) - J(u_*) = \int_\Omega \big\{F(Du(x)) - F(Du_*(x))\big\}\,dx \geq \int_\Omega \langle \zeta, D(u-u_*)(x)\rangle\,dx = 0,$$

using integration by parts (Theorem 20.7). □

We denote by $\mathrm{Lip}(k,\Omega)$ the set of functions $u : \Omega \to \mathbb{R}$ which are Lipschitz of rank k on Ω:

$$|u(x) - u(y)| \leqslant k\,|x-y| \;\; \forall x,y \in \Omega\,.$$

20.11 Proposition. *Let $\varphi : \Gamma \to \mathbb{R}$ satisfy a Lipschitz condition of rank K. Then, for any $k \geqslant K$, the problem of minimizing $J(u)$ over the functions $u \in \mathrm{Lip}(k,\Omega)$ which agree with φ on Γ has a solution.*

[1] The Hilbert-Haar approach extends to somewhat more general Lagrangians (see [6] and the references therein), but we shall not pursue this here.

Proof. We employ, naturally, the direct method. There is a function in $\mathrm{Lip}(k,\Omega)$ which extends φ to Ω (see Exer. 2.33). Since $k \geqslant K$, it follows that this function is admissible for the problem. Because Ω is bounded, $J(u)$ is well defined and finite for any admissible u. Thus, there is a minimizing sequence u_i for the problem P_k that consists of minimizing J relative to the functions in $\mathrm{Lip}(k,\Omega)$ which agree with φ on the boundary.

Since Du_i is a bounded sequence in $L^\infty(\Omega)^n$, and therefore in $L^2(\Omega)^n$, we may suppose (by taking subsequences) that each sequence $D_j u_i$ $(j = 1,2,\ldots,n)$ converges weakly in L^2 to a limit h_j. The u_i are equicontinuous, since they have a common Lipschitz constant k; they are also uniformly bounded on $\overline{\Omega}$. By Ascoli's theorem, we may further suppose (by taking a subsequence) that the sequence u_i converges uniformly to a limit u_*. It follows that $u_* \in \mathrm{Lip}(k,\Omega)$, and that u_* coincides with φ on Γ. Thus, u_* is admissible for P_k.

We prove next that $Du_* = h := (h_1,h_2,\ldots,h_n)$ a.e. Let $g \in \mathrm{Lip}_0(\Omega)$. Then, for any j, we have (see Theorem 20.7):

$$\begin{aligned}\int_\Omega D_j u_*(x)\,g(x)\,dx &= -\int_\Omega u_*(x)\,D_j g(x)\,dx = -\lim_{i\to\infty}\int_\Omega u_i(x)\,D_j g(x)\,dx \\ &= \lim_{i\to\infty}\int_\Omega D_j u_i(x)\,g(x)\,dx = \int_\Omega h_j(x)\,g(x)\,dx.\end{aligned}$$

Since this holds for any $g \in \mathrm{Lip}_0(\Omega)$, we deduce $Du_* = h$, as claimed.

It is now a direct consequence of Theorem 6.38 that

$$J(u_*) \leqslant \lim_{i\to\infty} J(u_i) = \inf P_k,$$

which confirms that u_* (which is admissible for P_k) is a solution of P_k. □

Minimizers. A function $u \in \mathrm{Lip}(k,\Omega)$ is called a **minimizer** for $\mathrm{Lip}(k,\Omega)$ if u minimizes J relative to the elements of $\mathrm{Lip}(k,\Omega)$ having the same boundary values as itself.

The following technical result will be of use.

20.12 Proposition. *Let u_1 and u_2 belong to $\mathrm{Lip}(k,\Omega)$, and define w by*

$$w(x) = \max\{u_1(x), u_2(x)\}.$$

Then $w \in \mathrm{Lip}(k,\Omega)$. If the derivatives $Dw(x)$, $Du_1(x)$, and $Du_2(x)$ exist at a point $x \in \Omega$ for which $u_1(x) = u_2(x)$, then they all coincide.

Proof. The fact that $w \in \mathrm{Lip}(k,\Omega)$ is long since familiar to the reader (see Exer. 2.32). To prove the remaining assertion, fix any $z \in \mathbb{R}^n$, and observe that, for any $t > 0$, we have

$$w(x+tz) - w(x) \geqslant u_1(x+tz) - u_1(z).$$

Dividing by t and letting t decrease to 0, we deduce $Dw(x) \bullet z \geqslant Du_1(x) \bullet z$. Since z is arbitrary, we find $Dw(x) = Du_1(x)$. The same argument applies to u_2, so the result follows. □

Since the set of points $x \in \Omega$ for which $Dw(x)$, $Du_1(x)$, and $Du_2(x)$ all exist is of full measure, we deduce:

20.13 Corollary. *We have* $Du_1 = Du_2$ a.e. *on the set* $\{x \in \Omega : u_1(x) = u_2(x)\}$.

20.14 Proposition. *Let u be a minimizer for Lip(k, Ω).*

(a) *If Ω' is an open subset of Ω, then $u|_{\Omega'}$ is a minimizer for Lip(k, Ω').*

(b) *Let $\tau \in \mathbb{R}^n$, and define $\Omega_\tau = \Omega + \tau$. Then the function $u_\tau(x) = u(x - \tau)$ is a minimizer for Lip(k, Ω_τ).*

(c) *For every constant c, the function $u + c$ is a minimizer for Lip(k, Ω).*

(d) *Any affine function belonging to Lip(k, Ω) is a minimizer for Lip(k, Ω).*

Proof. We prove (a), reasoning by the absurd. If the function $u|_{\Omega'}$ (which certainly belongs to $\text{Lip}(k, \Omega')$) is not a minimizer for $\text{Lip}(k, \Omega')$, then there exists an element u' in $\text{Lip}(k, \Omega')$ having the same values on $\partial\Omega'$ as u, and such that

$$\int_{\Omega'} F\big(Du'(x)\big)\,dx \;<\; \int_{\Omega'} F\big(Du(x)\big)\,dx.$$

Let us define

$$v(x) = \begin{cases} u'(x) & \text{for } x \in \Omega' \\ u(x) & \text{for } x \in \Omega \setminus \Omega'. \end{cases}$$

It is clear that v is continuous on $\overline{\Omega}$, and has the same values as u on Γ. Let us show that v is Lipschitz of rank k. In order to establish

$$|\,v(x) - v(y)\,| \leqslant k\,|\,x - y\,| \;\; \forall x, y \in \Omega\,,$$

it suffices to limit attention to the case $x \in \Omega \setminus \Omega'$, $y \in \Omega'$.

There exists a first $t \in [0,1]$ such that the point

$$z = (1-t)\,x + t\,y \in [x, y]$$

lies in $\partial\Omega'$. Then $u(z) = u'(z)$, and

$$\begin{aligned} |v(x) - v(y)| = |u(x) - u'(y)| &= |u(x) - u(z) + u'(z) - u'(y)| \\ &\leqslant k|x-z| + k|z-y| \;=\; k|x-y|, \end{aligned}$$

which confirms that $v \in \text{Lip}(k, \Omega)$.

In $\Omega \backslash \Omega'$, we have $Dv = Du$ a.e., by Cor. 20.13. In Ω', we have $Dv = Du'$. Thus

$$\int_{\Omega} \{ F(Dv(x)) - F(Du(x)) \} \, dx = \int_{\Omega'} \{ F(Du'(x)) - F(Du(x)) \} \, dx < 0.$$

This contradicts the fact that u is a minimizer for $\operatorname{Lip}(k, \Omega)$.

The last assertion of the theorem is a restatement of Prop. 20.10. The two remaining ones are left to the reader. □

20.15 Theorem. (The comparison principle) *Let F be strictly convex, and let u_1, u_2 be minimizers for $\operatorname{Lip}(k, \Omega)$ such that $u_1 \leqslant u_2$ on Γ. Then $u_1 \leqslant u_2$ in Ω.*

Proof. We define an element w in $\operatorname{Lip}(k, \Omega)$ as follows:

$$w(x) = \max\{ u_1(x), u_2(x) \}.$$

Note that $w = u_2$ on Γ. Since u_2 is a minimizer, we have

$$J(u_2) \leqslant J(w) = \int_{\{u_2 < u_1\}} F(Du_1(x)) \, dx + \int_{\{u_1 \leqslant u_2\}} F(Du_2(x)) \, dx,$$

by Cor. 20.13. We deduce

$$\int_{\{u_2 < u_1\}} F(Du_2(x)) \, dx \leqslant \int_{\{u_2 < u_1\}} F(Du_1(x)) \, dx.$$

An analogous argument using $\min\{u_1, u_2\}$ leads to the opposite inequality; thus, we have

$$\int_{\{u_2 < u_1\}} F(Du_2(x)) \, dx = \int_{\{u_2 < u_1\}} F(Du_1(x)) \, dx.$$

It follows that

$$J(u_2) = \int_{\Omega} F(Du_2(x)) \, dx = \int_{\Omega} F(Dw(x)) \, dx = J(w).$$

Now F is strictly convex, so if $Du_2 \neq Dw$ on a set of nonzero measure, we would have

$$J((u_2 + w)/2) < J(u_2)/2 + J(w)/2 = J(u_2).$$

This would contradict the fact that u_2 is a minimizer (since $(u_2 + w)/2$ coincides with u_2 on Γ). Thus Du_2 and Dw agree almost everywhere in Ω, which implies $w = u_2$ in Ω (by Exer. 10.30), which is precisely the required conclusion. □

20.16 Exercise. Let u_* be a minimizer for $\operatorname{Lip}(k, \Omega)$, where F is strictly convex. Prove that the maximum of u_* over the (compact) set $\overline{\Omega}$ is attained at some point in $\partial\Omega$. (This is sometimes referred to as the *maximum principle*.) □

The bounded slope condition. It turns out that the nature of the boundary condition can have substantial influence in determining the regularity of the solution. The following property will be central in this regard.

20.17 Definition. *We say that $\varphi : \Gamma \to \mathbb{R}$ satisfies the* **bounded slope condition** *with constant K if, for every point $\gamma \in \Gamma$, there exist two affine functions f^-, f^+ defined on $\mathbb{R}^n$ of the form*

$$f^-(y) = \langle \zeta_\gamma^-, y-\gamma \rangle + \varphi(\gamma), \quad f^+(y) = \langle \zeta_\gamma^+, y-\gamma \rangle + \varphi(\gamma)$$

(*thus, agreeing with φ at γ*) *with $|\zeta_\gamma^-| \leqslant K$, $|\zeta_\gamma^+| \leqslant K$, and such that*

$$f^-(y) \leqslant \varphi(y) \leqslant f^+(y) \;\forall\, y \in \Gamma.$$

The reader will observe that the bounded slope condition is a *joint* property of the function φ and the set Ω, a rather subtle one, as it turns out. The following makes certain observations about it.

20.18 Proposition.

(a) *If φ satisfies the bounded slope condition with constant K, then φ is Lipschitz of rank K on Γ.*

(b) *If φ coincides on Γ with an affine function, then φ satisfies the bounded slope condition, for a certain constant K.*

(c) *If φ does not coincide on Γ with an affine function, and if φ satisfies the bounded slope condition for a certain constant K, then Ω is convex.*

Proof. We leave the first two assertions as an exercise. For the third, let us consider any $\gamma \in \Gamma$. We see that the corresponding slopes ζ_γ^- and ζ_γ^+ in Def. 20.17 must differ, for otherwise φ would coincide on Γ with an affine function. The nonzero vector $\zeta = \zeta_\gamma^- - \zeta_\gamma^+$ then satisfies

$$\langle \zeta, y-\gamma \rangle \leqslant 0 \;\forall\, y \in \Gamma.$$

The same inequality then holds for all $y \in \operatorname{co}\Gamma$, which contains Ω since Ω is bounded (see Exer. 8.2). This shows that Ω admits a supporting hyperplane at each of its boundary points, and is therefore convex (Exer. 8.3). □

20.19 Exercise. Let $n = 2$, take Ω to be the unit ball in $\mathbb{R}^2$, and let φ be the function $\varphi(x) = x_1|x_1|^{1/2}$. Show that φ is continuously differentiable. Prove that the bounded slope condition fails to hold for any K. □

The following result has evolved from Hilbert's seminal work in 1904 on justifying the Dirichlet principle, with various mutations supplied over time by Hilbert, Haar,

Rado, von Neumann, Hartman, Nirenberg, and Stampacchia. It is generally named only for the first two, however, for evident reasons.

20.20 Theorem. (The Hilbert-Haar theorem) *We consider the basic problem relative to the class* $Lip(\Omega)$, *where the Lagrangian* $F(z)$ *is convex, and does not depend on* x *or* u. *If* φ *satisfies the bounded slope condition with constant* K, *then the problem admits a solution* u_* *which belongs to* $Lip(K,\Omega)$.

Proof. We first prove the theorem under a temporary additional hypothesis, whose removal will constitute the final step in the proof: F is strictly convex.

Choose $k > K$. Since φ is Lipschitz of rank K (see Exer. 20.18), Prop. 20.11 asserts the existence of $u_* \in \mathrm{Lip}(k,\Omega)$ which minimizes J over that set, subject to the boundary condition imposed by φ. We claim that u_* is Lipschitz on Ω of rank K, which turns out to be the main point of the proof.

To this end, fix any $x_0 \in \Omega$ and τ such that $x_0 - \tau \in \Omega$. We shall prove that

$$u_*(x_0 - \tau) \leqslant u_*(x_0) + K|\tau|, \tag{1}$$

which will establish the claim. Set $\Omega' = \Omega \cap (\Omega + \tau)$, an open subset of Ω containing x_0. We denote by u_1 the restriction to Ω' of the function $x \mapsto u_*(x-\tau)$, and by u_2 the restriction to Ω' of the function $x \mapsto u_*(x) + K|\tau|$. According to Prop. 20.14, both u_1 and u_2 are minimizers for $\mathrm{Lip}(k,\Omega')$. We proceed to show that $u_1 \leqslant u_2$ on $\partial\Omega'$. Let $y \in \partial\Omega'$. Then either y or $y-\tau$ belongs to $\partial\Omega$.

In the first case, by the bounded slope condition, there exists an affine function f, Lipschitz of rank K, such that $f(y) = \varphi(y) = u_*(y)$ and, on Γ, $u_* = \varphi \leqslant f$. Since f and u_* are minimizers for $\mathrm{Lip}(k,\Omega)$, the comparison principle 20.15 implies $u_* \leqslant f$ in Ω, whence

$$u_1(y) = u_*(y-\tau) \leqslant f(y-\tau) \leqslant f(y) + K|\tau| = u_*(y) + K|\tau| = u_2(y).$$

In the second case, there exists an affine function g, Lipschitz of rank K, such that $g(y-\tau) = \varphi(y-\tau) = u_*(y-\tau)$ and, on Γ, $u_* = \varphi \geqslant g$. Comparison yields $u_* \geqslant g$ in Ω, whence

$$u_1(y) = u_*(y-\tau) = g(y-\tau) \leqslant g(y) + K|\tau| \leqslant u_*(y) + K|\tau| = u_2(y).$$

In either case, then, we have $u_1 \leqslant u_2$ on $\partial\Omega'$. This implies $u_1 \leqslant u_2$ in Ω', by the comparison principle once more. In particular, $u_1(x_0) \leqslant u_2(x_0)$, which is precisely inequality (1), confirming the fact that u_* is Lipschitz on Ω of rank K.

We now prove that u_* solves the basic problem relative to $\mathrm{Lip}(\Omega)$. Let u in $\mathrm{Lip}(\Omega)$ be another admissible function. Then, for $\lambda \in (0,1)$ sufficiently small, the function $(1-\lambda)u_* + \lambda u$ is Lipschitz of rank less than k (since $K < k$), and equals φ on the boundary. Since u_* minimizes J over $\mathrm{Lip}(k,\Omega)$, we have

$$J(u_*) \leqslant J\big((1-\lambda)u_* + \lambda u\big) \leqslant (1-\lambda)J(u_*) + \lambda J(u).$$

We discover $J(u_*) \leqslant J(u)$. The theorem is proved, under the temporary additional hypothesis.

There remains to remove the hypothesis that F is strictly convex. Let $\varepsilon_i \downarrow 0$, and, for a fixed i, set $F_i(z) = F(z) + \varepsilon_i |z|^2$, which is strictly convex. We apply the strictly convex case of the theorem to obtain a function $u_i \in \text{Lip}(K, \Omega)$ which solves the basic problem (with Lagrangian F_i) relative to $\text{Lip}(\Omega)$.

As in the proof of Prop. 20.11, we may extract a subsequence (we do not relabel) converging uniformly to a limit $u_* \in \text{Lip}(K, \Omega)$ and such that

$$J(u_*) \leqslant \liminf_{i \to \infty} J(u_i).$$

Then, for any $u \in \text{Lip}(\Omega)$ having $u\big|_\Gamma = \varphi$, we have

$$\begin{aligned}
J(u_*) &\leqslant \liminf_{i\to\infty} \int_\Omega F\big(Du_i(x)\big)\,dx \\
&\leqslant \liminf_{i\to\infty} \int_\Omega \big\{F\big(Du_i(x)\big) + \varepsilon_i |Du_i(x)|^2\big\} dx = \liminf_{i\to\infty} \int_\Omega F_i\big(Du_i(x)\big)\,dx \\
&\leqslant \liminf_{i\to\infty} \int_\Omega F_i\big(Du(x)\big)\,dx = \liminf_{i\to\infty} \int_\Omega \big\{F\big(Du(x)\big) + \varepsilon_i |Du(x)|^2\big\} dx \\
&= \int_\Omega F\big(Du(x)\big)\,dx = J(u).
\end{aligned}$$

It follows that u_* solves the basic problem relative to $\text{Lip}(\Omega)$. □

We have seen in Prop. 20.18 that, except when φ is affine, the bounded slope condition can only hold when Ω is convex. Thus, in seeking to exploit the bounded slope condition (as we do below), there is little to lose by supposing *a priori* that Ω is convex.

We say that Ω is **uniformly strictly convex** if, in addition to being convex, Ω admits $c > 0$ such that, at every boundary point $\gamma \in \partial\Omega$, there is a hyperplane H passing through γ which satisfies:

$$d_H(y) \geqslant c|y - \gamma|^2 \;\; \forall\, y \in \Omega.$$

20.21 Proposition. *The open convex set Ω is uniformly strictly convex (with constant c) if and only if any point $\gamma \in \partial\Omega$ admits a unit normal vector $\nu \in N_{\text{cl}\,\Omega}(\gamma)$ such that*

$$\langle \nu, x - \gamma \rangle \leqslant -c|x - \gamma|^2 \;\; \forall\, x \in \Omega.$$

Proof. The following geometrical fact will be useful:

Lemma. *Let ν be a unit vector, let $\gamma \in \partial\Omega$, and let H be the hyperplane*

$$\big\{x \in \mathbb{R}^n : \langle \nu, x - \gamma \rangle = 0\big\}.$$

Suppose that $\langle \nu, x-\gamma \rangle < 0 \ \forall x \in \Omega$. *Then*

$$\operatorname{proj}_H(x) = x + d_H(x)\nu \ \ \forall x \in \Omega.$$

The hypotheses of the lemma evidently imply $H \cap \Omega = \emptyset$; fix any $x \in \Omega$, and set $p = \operatorname{proj}_H(x) \neq x$. Then p minimizes the function $y \mapsto |y-x|$ over the points $y \in H$. The corresponding necessary condition is $(p-x)/|p-x| \in -N_H(p)$, which implies that $p-x = r\nu$ for some scalar r (since $N_H = \mathbb{R}\nu$). We find that $r > 0$, as follows:

$$r = \langle p-x, \nu \rangle = \langle p-\gamma, \nu \rangle + \langle \gamma - x, \nu \rangle = 0 + \langle \gamma - x, \nu \rangle > 0.$$

Then $r = |r| = |p-x| = d_H(x)$, and the lemma is proved.

Now suppose that Ω is uniformly strictly convex with constant c. Fix any $\gamma \in \partial\Omega$, and let the hyperplane H provided by the definition of uniform strict convexity be given by $\{x \in \mathbb{R}^n : \langle \nu, x-\gamma \rangle = 0\}$, where ν is a unit vector. Then

$$\langle \nu, x-\gamma \rangle \neq 0 \ \ \forall x \in \Omega.$$

Since the image under a linear functional of a convex set is convex, it follows that $\langle \nu, x-\gamma \rangle$ is either strictly positive, or else strictly negative, on Ω. We choose ν in order to have the latter, which implies $\nu \in N_{\operatorname{cl}\Omega}(\gamma)$ (by the definition of a normal vector in the convex case).

Let x be any point in Ω, and set $p = \operatorname{proj}_H(x)$. Then, by the lemma, we have

$$\langle \nu, x-\gamma \rangle = \langle \nu, p - d_H(x)\nu - \gamma \rangle = -d_H(x) + \langle \nu, p-\gamma \rangle \leqslant -d_H(x) \leqslant -c|x-\gamma|^2.$$

This proves the "only if" implication of the proposition; we turn now to the converse.

Let us postulate the property relating to normal vectors; we wish to establish uniform strict convexity. Fix any $\gamma \in \partial\Omega$, and consider the hyperplane

$$H = \big\{ x : \langle \nu, x-\gamma \rangle = 0 \big\},$$

where ν is the unit normal vector corresponding to γ. Let x be any point in Ω, and set $p = \operatorname{proj}_H(x)$. Then we calculate

$$d_H(x) = |p-x| \geqslant \langle p-x, \nu \rangle = \langle p-\gamma, \nu \rangle + \langle \gamma - x, \nu \rangle = \langle \gamma - x, \nu \rangle \geqslant c|x-\gamma|^2,$$

which establishes uniform strict convexity. □

20.22 Exercise. Show that the unit ball in $\mathbb{R}^n$ is uniformly strictly convex, and that the unit square is not. □

It turns out that when the boundary of Ω is "uniformly curved" in the sense of uniform strict convexity, then *second* order smoothness of φ is enough to imply the bounded slope condition. (Exer. 20.19 shows that continuous differentiability of φ does not suffice, however.)

20.23 Theorem. (Miranda) *If Ω is bounded and uniformly strictly convex, then any function $\varphi : \Gamma \to \mathbb{R}$ which is the restriction to Γ of a function in $C^2(\mathbb{R}^n)$ satisfies the bounded slope condition (for a certain constant K depending on φ).*

Proof. We define $k = \max\{|\nabla\varphi(x)| : x \in \Gamma\}$ together with

$$L = \max\left\{-\langle w, \nabla^2\varphi(x)w\rangle/(2c) : |w| \leqslant 1, x \in \overline{\Omega}\right\}$$

and $K = L + k$, and we proceed to prove that φ satisfies the bounded slope condition with constant K.

Fix any $\gamma \in \Gamma$, and let $\nu \in N_{\mathrm{cl}\,\Omega}(\gamma)$ be associated with γ as in Prop. 20.21. Define

$$\zeta = \nabla\varphi(\gamma) + L\nu, \quad f(x) = \langle \zeta, x - \gamma\rangle + \varphi(\gamma).$$

Note that $|\zeta| \leqslant K$. Now let y be any point in Γ. The Taylor-Lagrange expansion yields a point z in $\overline{\Omega}$ such that

$$\varphi(y) - \varphi(\gamma) - \langle\nabla\varphi(\gamma), y - \gamma\rangle = \frac{1}{2}\langle(y-\gamma)\nabla^2\varphi(z), y - \gamma\rangle \geqslant -cL|y-\gamma|^2.$$

We calculate

$$\begin{aligned}
\varphi(y) - f(y) &= \varphi(y) - \langle\zeta, y-\gamma\rangle - \varphi(\gamma) \\
&= \varphi(y) - \varphi(\gamma) - \langle\nabla\varphi(\gamma), y-\gamma\rangle - L\langle\nu, y-\gamma\rangle \\
&\geqslant -cL|y-\gamma|^2 + cL|y-\gamma|^2 = 0.
\end{aligned}$$

Thus, f provides the function f^- in Def. 20.17. The proof of the existence of f^+ is similar; we omit it. □

We remark that the bounded slope condition forces φ to be affine on the 'flat parts' of Γ; see Exer. 21.40.

Together, the results proved above give a justification in Lipschitz terms of the Dirichlet principle (see Exer. 20.9):

20.24 Corollary. *Let $\Omega \subset \mathbb{R}^n$ be uniformly strictly convex, and let $\varphi : \mathbb{R}^n \to \mathbb{R}$ be of class C^2. Then there exists a unique $u_* \in Lip(\Omega)$ which agrees with φ on $\partial\Omega$ and which satisfies Laplace's equation in the weak sense:*

$$\int_\Omega \langle Du_*(x), D\psi(x)\rangle\,dx = 0 \;\; \forall\,\psi \in Lip_0(\Omega).$$

Proof. The bounded slope condition holds by Theorem 20.23, so that Theorem 20.20 is applicable. The resulting solution u_* satisfies Laplace's equation (in the weak sense) by Theorem 20.6 and, as seen in Exer. 20.9, corresponds to the unique weak solution. □

We remark that the function u_* of the corollary turns out to be smooth in the open set Ω: it can be shown to be harmonic there (a result known as *Weyl's lemma*). The fact that it is Lipschitz on Ω is an independent conclusion, of course.

20.25 Exercise. Formulate an existence result for the problem of Plateau (Example 20.4) based upon Theorems 20.20 and 20.23. □

20.4 Solutions in Sobolev space

The previous section dealt with Lagrangians depending only on Du. More general problems will fail, in general, to have solutions in $\mathrm{Lip}(\Omega)$. In the single-dimensional case ($n = 1$), the problem was extended to the class of absolutely continuous functions in order to develop an existence theory. As we have seen (p. 79), the generalization of this class to $n > 1$ is provided by the Sobolev spaces $W^{1,p}(\Omega)$, which are indeed the context for general existence theorems in the multiple integral calculus of variations. However, the theory becomes much more complex, for a variety of reasons, notably the regrettable fact that Sobolev functions may not be continuous (that is, may not admit a continuous representative). The following example of this phenomenon is a standard one that can be verified by elementary calculations.

20.26 Proposition. *Let $\alpha > 0$ and let $\Omega = B^\circ(0,1)$, the open unit ball in $\mathbb{R}^n$.*

(a) *The function $u(x) = |x|^{-\alpha}$ belongs to $L^1(\Omega)$ if and only if $\alpha < n$.*

(b) *The function u belongs to $W^{1,p}(\Omega)$ if and only if $\alpha < (n-p)/p$.*

It follows from the above that when $n > 1$, then $W^{1,1}(\Omega)$ contains discontinuous (and not locally bounded) functions.

20.27 Exercise. Prove that any $u \in \mathrm{Lip}(\Omega)$ belongs to $W^{1,\infty}(\Omega)$.[2] Deduce that u belongs to $W^{1,p}(\Omega)$ for every $p \geqslant 1$. (Recall: Ω is bounded by assumption.) □

We have seen (Exer. 5.10) that $W^{1,p}(\Omega)$ is a Banach space. We now verify that (for the expected values of p) it has the important property of being reflexive.

20.28 Theorem. *$W^{1,p}(\Omega)$ is reflexive if $1 < p < \infty$.*

---

[2] It can be shown that the converse is true when the boundary of Ω is "reasonable."

Proof. Consider the following subspace L of the space $X = L^p(\Omega)\times\big[L^p(\Omega)\big]^n$:

$$L = \big\{(u, Du) \in X : u \in W^{1,p}(\Omega)\big\}.$$

Note that the space X is a reflexive Banach space, as the product of such spaces. It follows from Exer. 5.8 that L is closed; thus, L is a reflexive Banach space (Exer. 5.49). We define a mapping $T : W^{1,p}(\Omega) \to L$ via $Tu = (u, Du)$. Note that T is an isometry between $W^{1,p}(\Omega)$ and L, for the right choice of (equivalent) product norms on X. Thus $W^{1,p}(\Omega)$ is revealed as being isometric to a reflexive Banach space, so that it inherits reflexivity by Prop. 5.42. □

It follows that $H^1(\Omega) := W^{1,2}(\Omega)$ is a Hilbert space with inner product

$$\langle u, v\rangle_{H^1(\Omega)} = \langle u, v\rangle_{L^2(\Omega)} + \textstyle\sum_{i=1}^{n} \langle D_i u, D_i v\rangle_{L^2(\Omega)}.$$

Boundary conditions. The possible discontinuity of Sobolev functions makes the issue of prescribing boundary conditions on Γ a nontrivial one in itself. When u and φ are continuous functions (as in the classical or the Lipschitz case of the basic problem), the meaning of "$u = \varphi$ on Γ" is clear. If u belongs to a Sobolev space $W^{1,p}(\Omega)$, however, then we may redefine u to equal φ on Γ, and we obtain the "same" u (that is, the same equivalence class). Clearly, then, *pointwise* equality cannot provide a meaningful way to specify the boundary condition in a Sobolev setting.

The same issue for the space $L^p(\Omega)$ would have no satisfactory resolution: there is no way to reasonably speak of the value on $\partial\Omega$ of a function $u \in L^p(\Omega)$ (see Exer. 21.36). Sobolev functions, however, are much more continuous (so to speak) than L^p functions. (Their discontinuities are few and averageable, a vague assertion that we have no intention of making precise.) It turns out that there is a natural (and unique) way to assign to $u \in W^{1,p}(\Omega)$ its value $\operatorname{tr} u$ on $\partial\Omega$, where $\operatorname{tr} u$, the *trace* of u, is a function in $L^p(\Gamma)$.

We adopt here a less technical route to specifying boundary values, one that is adequate for our purposes. It consists of defining an appropriate subspace of $W^{1,p}(\Omega)$. Recall that $C_c^\infty(\Omega)$ is the space of functions in $C^\infty(\Omega)$ having compact support in Ω, and which are therefore zero near the boundary.

20.29 Definition. *Let* $1 \leqslant p < \infty$. *We define*

$$W_0^{1,p}(\Omega) = \Big\{\lim_{i\to\infty} u_i : u_i \in C_c^\infty(\Omega)\Big\},$$

where the limit is taken in $W^{1,p}(\Omega)$ *(thus, with respect to the Sobolev norm).*

To put it another way, $W_0^{1,p}(\Omega)$ is the closure of $C_c^\infty(\Omega)$ in $W^{1,p}(\Omega)$. It follows that $W_0^{1,p}(\Omega)$ is a Banach space.

We interpret this new space as consisting of those elements in $W^{1,p}(\Omega)$ which are zero at the boundary. That this interpretation is reasonable depends on several non-trivial facts (not proved here), notably the following: when Γ is a C^1 manifold, and when u belongs to both $W^{1,p}(\Omega)$ and $C(\overline{\Omega})$, then

$$u = 0 \text{ on } \Gamma \iff u \in W_0^{1,p}(\Omega).$$

Armed with this concept of how to assert that a Sobolev function is zero on the boundary, we can formulate the basic problem relative to $W^{1,p}(\Omega)$:

$$\text{minimize } \ J(u) = \int_\Omega F\big(x, u(x), Du(x)\big)\, dx$$

relative to the functions $u \in W^{1,p}(\Omega)$ which satisfy $u - \varphi \in W_0^{1,p}(\Omega)$.

The reader will notice how the boundary condition is expressed by saying that the difference $u - \varphi$ vanishes at the boundary. Evidently, this formulation of the basic problem forces φ to belong to $W^{1,p}(\Omega)$ (in fact, we shall take $\varphi \in \text{Lip}(\Omega)$). Another point to retain is that the term Du in the integral now refers to the weak derivative of u (see p. 78).

Existence of a minimum. We now consider the existence question.

20.30 Theorem. *Let $\varphi \in Lip(\Omega)$. Suppose that the Lagrangian $F(x,u,z)$ is continuous in (x,u,z), convex with respect to (u,z), and coercive in the following sense: for certain constants $\alpha > 0$, $r > 1$ and β, for some function $\gamma \in L^{r_*}(\Omega)$ (where r_* is the conjugate exponent to r), we have*

$$F(x,u,z) \geqslant \alpha\big\{ |u|^r + |z|^r \big\} + \beta - \gamma(x)|u| \ \ \forall (x,u,z) \in \Omega \times \mathbb{R} \times \mathbb{R}^n.$$

Then the basic problem relative to $W^{1,r}(\Omega)$ admits a solution.

Proof. The hypotheses imply that the functional J is well defined and convex on $X = W^{1,r}(\Omega)$. It follows from Fatou's lemma that J is lower semicontinuous, relative to the norm topology of X. In view of the estimate (Hölder's inequality)

$$\int_\Omega |\gamma(x)|\,|u(x)|\, dx \leqslant \|\gamma\|_{L^{r_*}(\Omega)} \|u\|_{L^r(\Omega)},$$

the pointwise coercivity of F gives rise to the functional coercivity

$$J(u) \geqslant \widetilde{\alpha}\big\{ \|u\|^r_{L^r(\Omega)} + \|Du\|^r_{L^r(\Omega)} \big\} + \widetilde{\beta} = \widetilde{\alpha}\, \|u\|^r_{W^{1,r}(\Omega)} + \widetilde{\beta},$$

for certain constants $\widetilde{\alpha} > 0$ and $\widetilde{\beta}$. Since X is reflexive (Theorem 20.28), we may now apply directly the basic existence theorem 5.51. □

The reader might have expected, based upon the single integral case, that an existence theorem such as the above would hold *without* coercivity in the u variable. This expectation is quite justified, in fact:

20.31 Theorem. Theorem 20.30 *continues to hold if the coercivity is weakened to*

$$F(x,u,z) \geqslant \alpha|z|^r + \beta - \gamma(x)|u| \;\; \forall\,(x,u,z) \in \Omega\times\mathbb{R}\times\mathbb{R}^n.$$

Modifying the proof of Theorem 20.30 requires some Sobolev tools that we don't possess. Exer. 21.39 identifies these missing items, and shows how they can be used, together with the direct method, to prove Theorem 20.31.

The weak Euler equation. When a solution to the basic problem in $W^{1,1}(\Omega)$ exists, it is problematic to write the necessary conditions. The reader is already aware of this fact from the single integral theory: technical difficulties arise when Du is unbounded. One remedy is familiar, and works here too: we postulate additional structural assumptions of Tonelli-Morrey type, as in Theorem 16.13.

20.32 Theorem. *Let the Lagrangian $F(x,u,z)$ be measurable in x and locally Lipschitz in (u,z). Suppose that, for some constant c and function $d \in L^1(\Omega)$, for almost every $x \in \Omega$, we have*

$$\big|\big(F_u(x,u,z), F_z(x,u,z)\big)\big| \leqslant c\big\{|u|+|z|+\big|F(x,u,z)\big|\big\} + d(x)$$

for almost all (u,z) where the derivatives exist. Then, if u_* solves the basic problem relative to $W^{1,1}(\Omega)$, there exist summable functions $p : \Omega \to \mathbb{R}^n$ and $q : \Omega \to \mathbb{R}$ such that

$$\big(q(x), p(x)\big) \in \partial_C F\big(x, u_*(x), Du_*(x)\big) \text{ a.e. } x \in \Omega, \tag{E1}$$

and such that

$$\int_\Omega q(x)\,\psi(x) + \langle p(x), D\psi(x)\rangle\, dx = 0 \;\; \forall\,\psi \in Lip_0(\Omega). \tag{E2}$$

We remark that the conclusion is the same weak Euler equation (E1) (E2) asserted in the Lipschitz setting by Theorem 20.6. As before, the generalized gradient $\partial_C F$ appearing above is taken with respect to the (u,z) variables.

Proof. Fix $\psi \in \mathrm{Lip}_0(\Omega)$, where $|(\psi(x), D\psi(x))| \leqslant 1$ a.e., and define

$$g(x,s) = F\big(x, u_*(x)+s\psi(x), Du_*(x)+sD\psi(x)\big) - F(*),\;\; s \in [0,1],$$

as in the proof of Theorem 16.13, where $(*)$ denotes evaluation at $(x, u_*(x), Du_*(x))$. From the gradient formula (Theorem 10.27), we deduce that, for almost every x, we have, for all (u,z):

$$|(\alpha,\beta)| \leqslant 1 \implies F^\circ(x,u,z;(\alpha,\beta)) \leqslant c\big\{|u|+|z|+\big|F(x,u,z)\big|\big\} + d(x).$$

Whenever $g_s(x,s)$ exists, this implies that $|g_s(x,s)|$ is bounded above by

$$c\big\{2+|u_*(x)|+|Du_*(x)|+|F(*)|+|g(x,s)|\big\}+d(x).$$

Then the argument given in the proof of Theorem 16.13 allows us to bound g in such a way that we may invoke the dominated convergence theorem to derive

$$\int_\Omega F^\circ\big(*;(\psi,D\psi)\big)\,dx \geqslant 0.$$

Following this, the argument given in the proof of Theorem 20.6 leads to the required conclusion. □

Sufficiency of the weak Euler equation. The reader has come to expect our necessary conditions to become sufficient in a suitably convex setting. With due attention to integrability issues, this is the case in the Sobolev setting as well.

20.33 Theorem. *We suppose that the Lagrangian $F(x,u,z)$ is measurable in x and convex with respect to (u,z). Let $u_* \in W^{1,r}(\Omega)$ $(1 \leqslant r \leqslant \infty)$ be such that $J(u_*)$ is defined and finite. If there exist p and q in $L^{r_*}(\Omega)$ satisfying*

$$\big(q(x),p(x)\big) \in \partial_C F\big(x,u_*(x),Du_*(x)\big) \text{ a.e. } x \in \Omega \textit{ and}$$

$$\int_\Omega q(x)\,\psi(x)+\langle p(x),D\psi(x)\rangle\,dx = 0 \ \forall\,\psi \in C_c^\infty(\Omega),$$

then

$$J(u) \geqslant J(u_*) \ \forall\, u \in u_*+W_0^{1,r}(\Omega).$$

Proof. The hypotheses imply that the linear functional

$$\psi \mapsto \int_\Omega \big\{q(x)\,\psi(x)+\langle p(x),D\psi(x)\rangle\big\}\,dx$$

is defined and continuous on $W^{1,r}(\Omega)$. Fix any $\psi \in W_0^{1,r}(\Omega)$. Since this space is defined as the closure of $C_c^\infty(\Omega)$, we deduce

$$\int_\Omega \big\{q(x)\,\psi(x)+\langle p(x),D\psi(x)\rangle\big\}\,dx = 0.$$

The integrand on the left is bounded above almost everywhere by the generalized directional derivative $F^\circ(*;(\psi,D\psi))$ (see Def. 10.3), which coincides (since F is convex in (u,z), see Theorem 10.8) with the directional derivative $F'(*;(\psi,D\psi))$. But we have

$$F'\big(*;(\psi,D\psi)\big) \leqslant F\big(x,u_*+\psi,Du_*+D\psi\big)-F\big(x,u_*,Du_*\big),$$

by Prop. 2.22. Upon integrating, we discover $0 \leqslant J(u_*+\psi)-J(u_*)$. Since ψ is arbitrary in $W_0^{1,r}(\Omega)$, the conclusion follows. □

The next result shows, in particular, that the Lipschitz solutions of the basic problem provided by the Hilbert-Haar method (Theorem 20.20) are *also* solutions of the Sobolev version of the problem. In more general contexts, this can fail: the infimum of J (for given boundary conditions) over $W^{1,r}(\Omega)$ may be strictly less than over $\mathrm{Lip}(\Omega)$ (another instance of the Lavrentiev phenomenon).

20.34 Corollary. *Let $\varphi \in Lip(\Omega)$, and let the Lagrangian $F(z)$ be convex. Then if u_* solves the basic problem relative to $Lip(\Omega)$, it also solves it relative to $W^{1,r}(\Omega)$, for any $r \in [1,\infty)$.*

Proof. Fix any $r \in [1,\infty)$, and let p and q be the functions provided by Theorem 20.6; in the present setting, these functions are bounded. Then p and q lie in $L^{r_*}(\Omega)$, so the conclusion follows from Theorem 20.33. ☐

20.35 Exercise. Consider the following boundary value problem (DV), a variant of the Dirichlet problem studied earlier:

$$-\triangle u(x) + u(x) = f(x) \;\; (x \in \Omega), \quad u|_{\Gamma} = 0, \tag{DV}$$

where $f \in L^2(\Omega)$ is given. The space $H_0^1(\Omega) = W_0^{1,2}(\Omega)$ is the natural choice in which to consider weak solutions. A weak solution of (DV) in the $H_0^1(\Omega)$ sense refers to a function u in $H_0^1(\Omega)$ such that

$$\int_{\Omega} \{ \langle Du(x), D\psi(x) \rangle + u(x)\, \psi(x) \}\, dx = \int_{\Omega} f(x)\, \psi(x)\, dx \quad \forall\, \psi \in H_0^1(\Omega). \tag{1}$$

One can show that if a classical solution in $C^2(\overline{\Omega})$ of (DV) exists, then it is also a weak solution. The goal here is to prove

Theorem. *The functional*

$$J(u) = \int_{\Omega} \left\{ \tfrac{1}{2} \left| Du(x) \right|^2 + \tfrac{1}{2} u(x)^2 - f(x)\, u(x) \right\} dx$$

defined on $H_0^1(\Omega)$ attains a minimum at a unique point u which is the unique weak solution of (DV).

The proof is carried out in the following steps.

(a) Invoke Theorem 20.30 to deduce that J attains a minimum over $H_0^1(\Omega)$.

(b) Why is the minimizer u_* unique?

(c) Prove that u_* satisfies the weak Euler equation of Theorem 20.32.

(d) Prove that u_* is a weak solution of (DV) in the $H_0^1(\Omega)$ sense.

(e) Now let u be any weak solution of (DV). Invoke Theorem 20.33 to prove that u minimizes J over $H_0^1(\Omega)$, so that $u = u_*$. ☐

Once we know that a weak solution to a boundary-value problem such as (DV) exists, it is natural to ask whether it has greater regularity than merely that of a function in a Sobolev space. This topic is an important (and delicate) one in partial differential equations, one that we do not address.

20.36 Exercise. Hilbert space methods can also be used to study the boundary value problem (DV) of the preceding exercise, as we now see, under the same hypotheses.

(a) Prove that the map

$$u \mapsto \int_{\Omega} f(x)\, u(x)\, dx$$

defines a continuous linear functional on $H_0^1(\Omega)$.

(b) Invoke Theorem 7.2 to establish the existence of $u \in H_0^1(\Omega)$ satisfying (1).

Thus, we obtain the existence of a weak solution of (DV); uniqueness is easy to prove. An alternative route is the following:

(c) Obtain the theorem of the preceding exercise as a special case of the Lax-Milgram theorem (see Exer. 7.19). □

The use of Hilbert space methods as described in the exercise is effective when the underlying differential equation is *linear*, so that the associated Lagrangian is quadratic. The variational method, however, extends to other cases as well, as we now illustrate.

20.37 Example. We consider the following boundary value problem (D′):

$$\triangle u(x) = \begin{cases} +1 & \text{if } u(x) > 0 \\ -1 & \text{if } u(x) < 0 \\ \in [-1,1] & \text{if } u(x) = 0 \end{cases} \quad \big(x \in \Omega\big), \;\; u|_{\Gamma} = \varphi. \tag{D$'$}$$

We remark that the "Laplace inclusion" above is often expressed in the notation

$$\triangle u(x) = \operatorname{sgn} u(x),$$

where sgn is the signum function. A function $u \in H^1(\Omega)$ is a weak solution of (D′) if $u - \varphi \in H_0^1(\Omega)$, and if there exists a measurable function q satisfying

$$\int_{\Omega} q(x)\, \psi(x) + \langle Du(x), D\psi(x)\rangle\, dx = 0 \;\; \forall\, \psi \in H_0^1(\Omega) \tag{2}$$

as well as, for almost every $x \in \Omega$:

$$q(x) = \begin{cases} +1 & \text{if } u(x) > 0 \\ -1 & \text{if } u(x) < 0 \\ \in [-1,1] & \text{if } u(x) = 0. \end{cases} \tag{3}$$

We proceed to prove that if $\varphi \in \mathrm{Lip}(\Omega)$, then a unique such u exists.

The first step involves the minimization over $H^1(\Omega)$ of the functional

$$\int_\Omega \left\{ \tfrac{1}{2}\,|Du(x)|^2 + |u(x)| \right\} dx$$

subject to $u - \varphi \in H^1_0(\Omega)$. It is clear that we may apply Theorem 20.31, with $r = 2$; thus, a minimizing u_* exists. The convexity of the Lagrangian, which is strict in Du, can be shown to imply its uniqueness as a minimizer. We leave this as an exercise, along with a hint: use Poincaré's inequality (see Exer. 21.39).

Next, we wish to show that u_* is a weak solution of (D′). It is not difficult to verify that the Lagrangian of the problem satisfies the hypotheses of Theorem 20.32. Then, the function q provided by that theorem satisfies (3), and $p(x)$ reduces to $Du_*(x)$. It follows that (2) holds for every $\psi \in C^\infty_c(\Omega)$. Invoking density, it therefore holds for every $\psi \in H^1_0(\Omega)$. This confirms that u_* is a weak solution of (D′).

Finally, we observe that any other weak solution of (D′) is a minimizer for the problem above, by Theorem 20.33; it must therefore coincide with u_*. □

Chapter 21
Additional exercises for Part III

21.1 Exercise. Consider the problem

$$\min \int_1^3 \left\{ t\left(x'(t)\right)^2 - x(t) \right\} dt \,:\, x \in C^2[1,3],\ x(1) = 0,\ x(3) = -1.$$

(a) Find the unique admissible extremal x_*.

(b) Prove that x_* is a global minimizer for the problem.

(c) Prove that the problem

$$\min\ J(x) = \int_{-2}^3 \left\{ t\left(x'(t)\right)^2 - x(t) \right\} dt \,:\, x \in C^2[-2,3],\ x(-2) = A,\ x(3) = B$$

admits no local minimizer, regardless of the values of A and B. □

21.2 Exercise. Let $\Lambda(x,v) = v^2(1+v)^2$, and consider the problem

$$\min \int_0^1 \Lambda\left(x'(t)\right) dt \,:\, x \in C^2[0,1],\ x(0) = 0,\ x(1) = m.$$

(a) Is Λ convex?

(b) Show that $x_*(t) = t\,m$ is an admissible extremal.

(c) Show that when $m = -1/6$, x_* is a weak local minimizer.

(d) Show that when $m = -1/2$, x_* is a weak local maximizer.

(e) Show that when $m = -1/6$, x_* is not a global minimizer. □

21.3 Exercise. Consider the following problem:

$$\min \int_0^1 \left\{ x'(t)^2 + 2t^2 x(t) + x(t)^2 \right\} dt \,:\, x \in \mathrm{Lip}[0,1],\ x(0) = 0,\ x(1) = 1.$$

F. Clarke, *Functional Analysis, Calculus of Variations and Optimal Control*,
Graduate Texts in Mathematics 264, DOI 10.1007/978-1-4471-4820-3_21,

(a) Find the unique admissible extremal x_*.

(b) Prove that x_* is a global minimizer.

(c) What arc x solves the problem if the constraint $x(1) = 1$ is removed? □

21.4 Exercise. We consider the problem

$$\min \int_0^1 e^{x(t)}\big(1+x'(t)^2\big)\,dt \;:\; x \in C^2[0,1],\; x(0)=0,\; x(1)=1.$$

(a) Show that the function $x_*(t) = t$ is an admissible extremal.

(b) Is the Lagrangian convex in (x,v)?

(c) Prove that x_* provides a weak local minimum relative to $C^2[0,1]$.

(d) Prove that x_* provides a global minimum. □

21.5 Exercise. (Queen Dido's problem) We consider arcs $x \in C^2[0,1]$ satisfying

$$x(t) \geqslant 0 \;\; \forall\, t \in [0,1], \;\; x(0) = x(1) = 0,$$

and such that the corresponding curve has prescribed length:

$$\int_0^1 \sqrt{1+x'(t)^2}\,dt \;=\; L > 1\,.$$

The classical problem of Dido is to find the arc x_* which maximizes the area under the curve, under the given constraints. Note that the area in question is given by

$$\int_0^1 x(t)\,dt.$$

If Theorem 14.21 applies here, which is not completely clear (because of the presence of the state constraint $x \geqslant 0$), we are led to consider the augmented Lagrangian

$$-x + \lambda\sqrt{1+v^2}\,.$$

The corresponding Euler equation is

$$\frac{d}{dt}\,\frac{\lambda\, x'}{\sqrt{1+x'^2}} \;=\; -1,$$

which implies $\lambda\, x'/\sqrt{1+x'^2} \;=\; -t+k$. Solving this separable differential equation leads to the conclusion

$$\big(x(t)-c\big)^2 + (t-k)^2 \;=\; \lambda^2,$$

for certain constants c and k. Thus, the curve x_* we are seeking describes, perhaps, a circular arc. By symmetry, the arc should probably be centered at a point of the form $(1/2, z)$. Observe that this conjecture is feasible only if $L < \pi/2$. We make this assumption, and we proceed to prove by an inductive argument that we have identified a curve that maximizes the area.

(a) Show that the circular arc corresponds to an extremal of the Lagrangian

$$\Lambda_+(x,v) = -x + \lambda\sqrt{1+v^2}$$

for a certain $\lambda > 0$.

(b) Observe that Λ_+ is convex, and use this to prove that x_* solves the problem

$$\min \int_0^1 \Lambda_+\big(x(t), x'(t)\big)\, dt \; : \; x \in C^2[0,1], \; x(0) = x(1) = 0.$$

(c) Conclude that x_* is the sought-for maximizing curve. □

21.6 Exercise. Among all curves joining a given point $(0, A)$ on the positive x-axis to some point $(b, 0)$ on the t-axis (where $b > 0$ is unspecified) and enclosing a given area S together with the t-axis, find the curve which generates the least area when rotated about the t-axis. □

21.7 Exercise. Find a Lagrangian $\Lambda(t, x, v)$ whose Euler equation is

$$a)\; x'' = x^3 \qquad b)\; x'' + x' - 1 = 0 \qquad c)\; x''x + x' = 1.$$ □

21.8 Exercise. Prove that for $T > 0$ sufficiently large, there exists a nontrivial solution $x(\cdot) \in C^\infty[0, T]$ of the boundary-value problem

$$x''(t) + \sin x(t) + \big(\sin x(t)\big)^3 = 0, \; t \in [0, T], \; x(0) = x(T) = 0.$$ □

21.9 Exercise. Consider the basic problem (P) with a Lagrangian $\Lambda \in C^3$. Let x_* in $C^2[a, b]$ be an admissible extremal satisfying the strengthened Legendre condition. Show that x_* admits no conjugate points in (a, b) if one of the following holds:

(a) $\Lambda(t, x, v)$ is of the form $\Lambda(t, v)$ (independent of x);

(b) $\Lambda(t, x, v)$ is convex in (x, v). □

21.10 Exercise. Find the solution of the following problem, or else prove that none exists:

$$\min \int_0^1 e^{x(t)} e^{x'(t)}\, dt \; : \; x \in \mathrm{Lip}[0,1], \; x(0) = 0, \; x(1) = 1.$$ □

21.11 Exercise. When the soap bubble problem (see Example 14.5) is restricted to the class of curves parametrizable in the form $t(x)$, it amounts to minimizing

$$J(t(\cdot)) = \int_A^B x\sqrt{1+t'(x)^2}\,dx.$$

Use this to strengthen the conclusion of Exer. 14.16 as follows: for any $T > 0$, the catenary $\cosh t$ is a *global* minimizer on $[0,T]$ relative to such curves. □

21.12 Exercise. Find all the weak local minima and maxima of the functional

$$\int_0^1 (1+t)\,x'(t)^2\,dt$$

over AC$[0,1]$, under the boundary conditions $x(0) = 0,\ x(1) = 1$. □

21.13 Exercise. (Approximation of Lipschitz arcs)

Let $h : [a,b] \to \mathbb{R}$ be a Lipschitz function. We prove:

Theorem. *For any $\varepsilon > 0$, there exists a polynomial $g(t)$ with $g(a) = h(a)$ and $g(b) = h(b)$ such that*

$$\|g'\| \leqslant \|h'\| + \varepsilon, \quad \|g-h\| + \|g'-h'\|_{L^1} < \varepsilon.$$

(As usual, the anonymous norm is that of L^∞.)

(a) Invoke Lusin's theorem (which is stated on p. 112) to find a continuous function f on $[a,b]$ such that

$$\|f\| \leqslant \|h'\|,\ f = h' \text{ except on a set of measure at most } \varepsilon.$$

(b) Use the Weierstrass approximation theorem to find a polynomial p for which $\|f - p\| < \varepsilon$.

(c) Set $\varphi(t) = h(a) + \int_a^t p(s)\,ds$. Show that, for a certain constant c satisfying

$$|c| \leqslant \varepsilon\{b-a+2\|h'\|\}/(b-a),$$

the function $g(t) = \varphi(t) + c(t-a)$ agrees with h at a and b and satisfies

$$\|g'\| \leqslant \|h'\| + \varepsilon + |c|,$$
$$\max\big[\|g-h\|, \|g'-h'\|_{L^1}\big] \leqslant \varepsilon\{b-a+2\|h'\|\} + (b-a)|c|.$$

Use these estimates to deduce the theorem (for a suitably redefined ε).

(d) We proceed to apply the theorem to the functional

$$J(x) = \int_a^b \Lambda\big(t, x(t), x'(t)\big)\,dt,$$

where Λ is locally Lipschitz. Prove that for any $x \in \operatorname{Lip}[a,b]$, for any $\varepsilon > 0$, there exists a polynomial y having the same values as x at a and b such that $\|x-y\| < \varepsilon$ and $|J(x)-J(y)| < \varepsilon$. (Such a result is referred to as a *fairing theorem*.)

(e) Consider now the problem

$$\min \; J(x) \, : \; x \in X, \; x(a)=A, \; x(b)=B, \; \|x\| < r, \; \|x'\| < R,$$

where X is a subset of $\operatorname{Lip}[a,b]$ and $r, R \in (0,\infty]$. Show that the infimum in the problem is the *same* for any X that contains all polynomials.

(f) Prove that if $x_* \in C^2[a,b]$ is a weak local minimizer for the basic problem (P) relative to $C^2[a,b]$, where Λ is locally Lipschitz, then it is also a weak local minimizer relative to $\operatorname{Lip}[a,b]$. Show that the same holds for a strong local minimizer. □

21.14 Exercise. (Wirtinger's inequality, conclusion) Prove the following inequality for any function $y \in \operatorname{Lip}[a,b]$ which vanishes at a and b:

$$\int_a^b y(t)^2\,dt \;\leqslant\; \frac{(b-a)^2}{\pi^2} \int_a^b y'(t)^2\,dt,$$

with equality if and only if $y(t) = c\sin[\pi(t-a)/(b-a)]$ for some constant c. □

21.15 Exercise. It follows from Exer. 14.17 that the infimum in the following problem is $-\infty$:

$$\min \int_0^{2\pi} \big(x'(t)^2 - x(t)^2\big)\,dt \; : \; x \in \operatorname{AC}[0,2\pi], \; x(0)=0.$$

Solve the problem when the auxiliary state constraint $0 \leqslant x(t) \leqslant 1$ is imposed. □

21.16 Exercise. We return to the problem of Example 16.1:

$$\min \int_0^1 \big(1+x(t)\big)\,x'(t)^2\,dt \; : \; x \in C^2[0,1], \; x(0)=0, \; x(1)=3.$$

(a) Show that the function $x_*(t) = (7t+1)^{2/3} - 1$ is a weak local minimizer. (We have seen that x_* is not a global minimum.)

(b) Suppose now that the auxiliary constraint $x(t) \geqslant 0$ is added to the problem. Show that x_* provides a global minimum in that case. □

21.17 Exercise. We study the solutions $x \in C^\infty[0,T]$ of the following boundary value problem:

$$x''(t) = x^3(t) + b\,x^2(t) + c\,x(t) + d\sin t, \; 0 \leqslant t \leqslant T, \; x(0)=x(T)=0, \qquad (1)$$

where $T > 0$ is given, as well as the parameters b, c, d.

(a) Formulate a problem (P) in the calculus of variations whose Euler equation is the differential equation above, and show with its help that there exists a solution x of (1).

(b) Prove that the solution of (1) is unique when $b^2 \leqslant 3c$.

The question of the *nontriviality* of the solution of (1) arises when $d = 0$, since in that case, $x \equiv 0$ is a solution. When $d = 0$ and $b^2 \leqslant 3c$, the unique solution of (1) is $x \equiv 0$, by the preceding.

(c) Prove that for $d = 0$ and $c < 0$, there is a nontrivial solution of (1) when the horizon T satisfies $T > \pi |c|^{-1/2}$. □

21.18 Exercise. We study the existence question for certain variational problems with auxiliary constraints.

(a) Consider the problem

$$\text{minimize} \int_a^b \Lambda\big(t, x(t), x'(t)\big)\, dt \;:\; x'(t) \in V \text{ a.e., } x(a) = A,\ x(b) = B,$$

where Λ is continuous in (t, x, v) and convex in v, and where V is a compact convex subset of $\mathbb{R}^n$. Prove that a solution $x_* \in \operatorname{Lip}[a,b]$ exists, provided there exists at least one admissible arc.

(b) Consider next the isoperimetric problem

$$\text{minimize} \int_a^b \Lambda\big(t, x(t), x'(t)\big)\, dt \;\text{ subject to}$$
$$\int_a^b \langle h(t, x(t)), x'(t)\rangle\, dt = c,\ x(a) = A,\ x(b) = B,$$

where Λ satisfies the hypotheses of Theorem 16.2, $h : [a,b] \times \mathbb{R}^n \to \mathbb{R}^n$ is continuous, and $c \in \mathbb{R}$ is given. Prove that a solution $x_* \in \operatorname{AC}[a,b]$ exists, provided there is at least one admissible arc.

(c) Consider now the problem

$$\text{minimize} \int_a^b \Lambda\big(t, x(t), x'(t)\big)\, dt \;\text{ subject to}$$
$$g\big(t, x(t), x'(t)\big) \leqslant 0 \text{ a.e., } x(a) = A,\ x(b) \in E,$$

where Λ and g are continuous in (t, x, v) and convex in v, E is closed, and where either Λ or g satisfies the coercivity hypothesis of Theorem 16.2. Prove that a solution $x_* \in \operatorname{AC}[a,b]$ exists, provided there is at least one admissible arc. □

21.19 Exercise. Solve the problem of Example 19.4 deductively; that is, by proving existence and applying appropriate necessary conditions. □

21.20 Exercise. Let $\ell : \mathbb{R} \to [0,\infty)$ be a continuous function. Prove that the following problem admits a solution:

$$\min\ \ell(x(1)) + \int_0^1 \{x(t)^3 + x'(t)^2\}\,dt\ :\ x \in \operatorname{Lip}[0,1],\ x'(t) \geqslant 0 \text{ a.e.},\ x(0) = 0.$$

□

21.21 Exercise. Consider the following problem (P) in the calculus of variations:

$$\min \quad \int_0^1 \Big\{\sqrt{1+x'(t)^2} - x(t)\Big\}\,dt\ :\ x \in \operatorname{AC}[0,1],\ x(0) = 1,\ x(1) = 0.$$

(a) Show that $x_*(t) = \sqrt{1-t^2}$ is an admissible function for the problem, one that fails to lie in $\operatorname{Lip}[0,1]$.

(b) Show that x_* satisfies the integral Euler equation.

(c) Prove that x_* is the unique solution of (P).

(d) Deduce that x_* is the unique solution of the isoperimetric problem

$$\min\ \int_0^1 \sqrt{1+x'(t)^2}\,dt\ :\ x \in \operatorname{AC}[0,1],\ \int_0^1 x(t)\,dt = \frac{\pi}{4},\ x(0) = 1,\ x(1) = 0.$$

(e) Prove that (P) admits no solution when the underlying space is taken to be $\operatorname{Lip}[0,1]$ rather than $\operatorname{AC}[0,1]$.

□

21.22 Exercise. Prove that the boundary-value problem

$$x''(t) = x(t)\sin x(t) - \cos x(t), \quad x(0) = 0,\ x(1) = 0$$

admits a solution $x \in C^\infty[0,1]$.

□

21.23 Exercise. Consider the problem depicted in Exer. 17.3.

(a) Suppose that $x_* \in C^4[a,b]$ satisfies the second-order Euler equation and the given boundary conditions. If Λ is convex with respect to (x,v,w), prove that x_* is a global solution of the problem.

(b) Prove that the problem of minimizing

$$\int_0^1 \{x''(t)^2 - 48\,x(t)\}\,dt$$

subject to

$$x \in \operatorname{AC}[0,1],\ x' \in \operatorname{AC}[0,1],\ x(0) = 0,\ x(1) = 1,\ x'(0) = 1,\ x'(1) = 1$$

has a solution, and identify it.

□

21.24 Exercise. Let $f : \mathbb{R}^n \to \mathbb{R}$ be locally Lipschitz and linearly bounded:

$$|f(x)| \leqslant c + d|x| \;\; \forall x \in \mathbb{R}^n.$$

Let A, B be given points in $\mathbb{R}^n$. Prove that there exists $x \in C^1[a,b]$ with x' in $AC[0,1]$ satisfying

$$x''(t) \in \partial_C f\big(x(t)\big),\; t \in [0,1] \text{ a.e.},\; x(0) = A,\; x(1) = B. \qquad \square$$

21.25 Exercise. Find a smooth function $\varphi : (0,\pi) \times \mathbb{R} \to \mathbb{R}$ such that

$$v^2 - x^2 \geqslant \varphi_t(t,x) + \varphi_x(t,v) \;\; \forall (t,v) \in (0,\pi) \times \mathbb{R}.$$

Use φ to give an elementary proof that, for any $x \in \mathrm{Lip}_0[0,\pi]$, we have

$$\int_0^\pi \big\{x'(t)^2 - x(t)^2\big\}\, dt \geqslant 0. \qquad \square$$

21.26 Exercise. The classical problem of Zenodoros consists of finding the solid of revolution of maximal volume having given surface area; the conjecture is that the solution is provided by a semicircle.[1] Using a substitution due to Euler, the problem can be reduced to proving that the arc $x_*(t) = 2t - 2t^2$ solves the following problem (P) in the calculus of variations:

$$\text{minimize } J(x) = -\int_0^1 \sqrt{x(t)\big(4 - x'(t)^2\big)}\, dt$$
$$\text{subject to } x(t) \geqslant 0,\; |x'(t)| \leqslant 2,\; x(0) = x(1) = 0.$$

Note that the problem includes both a state constraint and a differential constraint, and that admissible arcs necessarily satisfy $0 \leqslant x(t) \leqslant 2t$.

(a) Prove that $t \geqslant 0,\; 2t \geqslant x \geqslant 0 \implies 2(t-x)^2 + t^2v^2 + 4t(t-x)v \geqslant 0 \;\forall v$.

We obtain $(2t - xv + tv)^2 \geqslant x(4 - v^2)(2t - x)$, by rearranging. If we restrict attention to $|v| \leqslant 2$, then $2t - xv + tv \geqslant 0$, and we deduce:

$$2t - xv + tv \geqslant \sqrt{x\big(4 - v^2\big)(2t - x)}\,.$$

(b) Show that, for all $t \in (0,1],\; 0 \leqslant x < 2t,\; |v| \leqslant 2$, we have

$$\Lambda(t,x,v) \geqslant (-2t + xv - tv)/\sqrt{2t - x} = \varphi_t(t,x) + \varphi_x(t,x)\,v,$$

where Λ is the Lagrangian of the problem (P), and where

$$\varphi(t,x) = -(2/3)(t + x)\sqrt{2t - x}\,.$$

[1] See Troutman [39].

(c) Let x be any admissible arc for (P) that satisfies $x(t) < 2t$ for $t \in (0,1]$. Prove that $J(x) \geqslant -2\sqrt{2}/3$.

(d) Extend this conclusion to all admissible arcs. [Hint: replace x by λx with $\lambda < 1$, and invoke the preceding.]

(e) Show that the lower bound is attained for $x_*(t) = 2t - 2t^2$, which is therefore revealed as the solution to (P).

Can you explain how the very useful function φ was found? □

21.27 Exercise. (An indirect existence method) The purpose of this exercise is to illustrate an indirect approach to existence when the Lagrangian fails to possess the coercivity postulated in Tonelli's theorem.[2] We consider the following problem (P):

$$\min \; J(x) = \int_a^b g\big(x(t)\big)\sqrt{1+|x'(t)|^2}\,dt \; : \; x(a)=A\,,\, x(b)=B,\, x \in \operatorname{Lip}[a,b],$$

where $g : \mathbb{R}^n \to \mathbb{R}$ is continuous and bounded below by a positive constant δ. We establish the existence of a solution x_* to (P) belonging to $C^1[a,b]$.

The proof consists of examining a sequence of perturbed problems which are coercive and admit solutions, and showing that these solutions are regular, and converge appropriately to a solution of the original problem.

(a) Why does Tonelli's theorem not apply in this setting?

(b) Let ε_i be a positive sequence strictly decreasing to 0, and set

$$\Lambda_i(x,v) \;=\; g(x)\sqrt{1+|v|^2} + \varepsilon_i |v|^2.$$

For given i, prove the existence of a function x_i which solves

$$\min \; \int_a^b \Lambda_i\big(x(t), x'(t)\big)\,dt \; : \; x \in \operatorname{AC}[a,b],\; x(a) = A\,,\; x(b) = B.$$

(c) Show that x_i belongs to $C^1[a,b]$, and that there exists a constant λ_i such that

$$\frac{g(x_i(t))}{\sqrt{1+|x_i'(t)|^2}} - \varepsilon_i |x_i'(t)|^2 \;=\; \lambda_i\,,\;\; t \in [a,b] \text{ a.e.}$$

(d) Prove the existence of M such that

$$\int_a^b \Lambda_i\big(x_i(t), x_i'(t)\big)\,dt \;\leqslant\; M \;\; \forall\, i\,.$$

---

[2] See Clarke [14] for more general developments along this line.

Deduce from this the existence of a constant R such that $\|x_i\| \leqslant R$ for all i, and prove that

$$\min_{a \leqslant t \leqslant b} |x_i'(t)| \leqslant M/\big(\delta(b-a)\big) =: m \ \forall i.$$

(e) Show that $\lambda_i \geqslant \delta/(2\sqrt{1+m^2})$ for all i sufficiently large.

(f) Set $g_0 = \max\{g(x) : |x| \leqslant R\}$. Deduce that, for all i sufficiently large, we have

$$\max_{a \leqslant t \leqslant b} |x_i'(t)|^2 \leqslant 4\big(1+m^2\big) g_0^2/\delta^2 - 1.$$

(g) Prove the existence of a subsequence of x_i converging uniformly to a function $x_* \in \operatorname{Lip}[a,b]$ such that $J(x_*) \leqslant J(x)$ whenever x is admissible for (P).

(h) Prove that x_* lies in $C^1[a,b]$ and conclude. □

21.28 Exercise. (The Legendre transform) We suppose that $\Lambda : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ is a Lagrangian of class C^2 satisfying $\Lambda_{vv} > 0$ globally.

(a) Let $\bar{x}, \bar{v}, \bar{p}$ in $\mathbb{R}^n$ satisfy the equation $p = \Lambda_v(x,v)$. Prove that this equation defines a function $v(x,p)$ of class C^1 on a neighborhood of $(\bar{x},\bar{p})$ such that $v(\bar{x},\bar{p}) = \bar{v}$. We set

$$H(x,p) = \langle p, v \rangle - \Lambda(x,v),$$

where $v = v(x,p)$. H is the *Hamiltonian* obtained from Λ by applying the *Legendre transform*. Prove that H is of class C^1 in a neighborhood of $(\bar{x},\bar{p})$.

(b) Now let $x_* \in C^2[a,b]$, and define a function p_* and a set S by

$$p_*(t) = \Lambda_v\big(x_*(t), x_*'(t)\big), \quad S = \big\{ \big(x_*(t), p_*(t)\big) : t \in [a,b] \big\}.$$

Prove that there is a neighborhood V of S on which is defined a unique function v of class C^1 such that

$$p = \Lambda_v\big(x, v(x,p)\big) \ \forall (x,p) \in V, \quad v\big(x_*(t), p_*(t)\big) = x_*'(t) \ \forall t \in [a,b].$$

In this neighborhood V, we define H as above.

(c) Prove that x_* solves the Euler equation on $[a,b]$, that is, satisfies

$$\frac{d}{dt} \Lambda_v\big(x(t), x'(t)\big) = \Lambda_x\big(x(t), x'(t)\big)$$

if and only if (x_*, p_*) satisfies the *Hamiltonian system*

$$-p'(t) = H_x\big(x(t), p(t)\big), \quad x'(t) = H_p\big(x(t), p(t)\big).$$

If this is the case, show that H is constant along (x_*, p_*).

(d) We suppose now that Λ is coercive of degree $r > 1$ (see Theorem 16.2). Prove that H coincides with the Fenchel conjugate of Λ, in the sense that

$$H(x,p) \;=\; \max_{v\,\in\,\mathbb{R}^n} \,\{\,\langle\, p,v\,\rangle - \Lambda\,(x,v)\,\}\,.$$

Deduce that H is convex in the p variable. □

21.29 Exercise. (Duality) Let $\Lambda\,(x,v)$ be continuously differentiable, and set

$$L(p,q) \;=\; \inf\big\{\,\Lambda\,(x,v) - \langle\, p,v\,\rangle - \langle\, q,x\,\rangle \;:\; x,v \in \mathbb{R}^n\,\big\}.$$

We assume that L is finite-valued.

(a) Show that L is concave, and that

$$-\langle A,\, p(a)\rangle + \int_a^b L\big(p(t),\, p'(t)\big)\,dt \;\leqslant\; \langle \beta, x(b)\rangle + \int_a^b \Lambda\,\big(x(t), x'(t)\big)\,dt$$

for all $x, p \in C^1[\,a,b\,]$ satisfying $x(a) = A,\; p(b) = -\beta$. (Why are the integrals well defined?)

(b) Let (P) be the problem

$$\text{minimize } \;\langle\,\beta, x(b)\,\rangle + \int_a^b \Lambda\,\big(x(t), x'(t)\big)\,dt \;:\; x \in C^1[\,a,b\,],\; x(a) = A.$$

Let (D) be the problem

$$\text{maximize } \;-\langle\, A,\, p(a)\,\rangle + \int_a^b L\big(p(t),\, p'(t)\big)\,dt \;:\; p \in C^1[\,a,b\,],\; p(b) = -\beta.$$

Prove that $\sup(\text{D}) \leqslant \inf(\text{P})$.

(c) If (P) admits a solution x_*, and if Λ is convex, prove that $\max(\text{D}) = \min(\text{P})$, and that the costate $p_*(t) = \Lambda_v(x_*(t), x_*'(t))$ lies in $C^1[\,a,b\,]$ and solves (D).

(d) Formulate a set of hypotheses guaranteeing that (P) admits a solution. □

21.30 Exercise. (Periodic Hamiltonian trajectories 1) Let $H : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ be a given continuously differentiable function. We study the Hamiltonian system of equations

$$-\,p'(t) \;=\; H_x\big(x(t),\, p(t)\big),\;\; x'(t) \;=\; H_p\big(x(t),\, p(t)\big) \tag{2}$$

where x and p are arcs. More specifically, the problem we consider is the possible existence on some interval $[\,0,T\,]$ of a solution (x,p) of these equations satisfying $x(0) = x(T),\; p(0) = p(T)$, and also $H(x,p) = c$, where c is a given constant. (It is easy to see that any solution to (2) automatically lies on a level surface of H.) We call such a solution *periodic*. Thus we seek a periodic trajectory of (2) having *prescribed energy* c (that is, lying in $H^{-1}(c)$); the period T is unknown, however.

The existence of a periodic trajectory can only be guaranteed under suitable conditions on H. We shall use here a variational method known as the *dual action principle* [3] in order to obtain the following result.

Theorem. *Let $H^{-1}(c)$ be the boundary of a compact, strictly convex set containing 0 in its interior, and suppose $\nabla H \neq 0$ on $H^{-1}(c)$. Then, for some $T > 0$, there is a periodic solution of (2) on $H^{-1}(c)$.*

The reader is asked to prove each of the claims in the proof below.

Define $h(x,p)$ to be λ^2, where λ is the unique positive scalar such that $(x,p)/\lambda$ lies in $H^{-1}(c)$ (we set $h(0,0)=0$). We recognize h as the square of the gauge function corresponding to $H^{-1}(c)$ (see Theorem 2.36); evidently, we have $h^{-1}(1)=H^{-1}(c)$.

Claim:

1) h is C^1. [Hint: use the implicit function theorem.]

2) h is positively homogeneous of degree 2 :

$$h\big(\lambda\alpha,\lambda\beta\big) = \lambda^2 h\big(\alpha,\beta\big) \ \ \forall\lambda > 0, \ \forall(\alpha,\beta).$$

3) h is strictly convex, and, for certain positive constants δ, Δ, we have

$$\delta\big|(\alpha,\beta)\big|^2 \leqslant h(\alpha,\beta) \leqslant \Delta\big|(\alpha,\beta)\big|^2 \ \ \forall(\alpha,\beta)\in\mathbb{R}^n\times\mathbb{R}^n.$$

4) $|\nabla h|$ is bounded away from zero on $H^{-1}(c)$ and, for any (x,p) in $H^{-1}(c)$, the vector $\nabla h(x,p)$ is a nonzero multiple of $\nabla H(x,p)$.

Suppose now that we have a periodic solution (x,p) of (2) for H replaced by h, and with (x,p) lying in $h^{-1}(1)$. Let its period be T. Then, by the above, there is a bounded function λ on $[0,T]$, bounded away from 0, such that

$$-p'(t) = \lambda(t)H_x\big(x(t),p(t)\big), \ \ x'(t) = \lambda(t)H_p\big(x(t),p(t)\big), \ \ t\in[0,T] \text{ a.e.}$$

One may take λ to be measurable. The bi-Lipschitz time rescaling

$$\tau(t) = \int_0^t \lambda(s)\,ds$$

gives rise to functions

$$\big(\tilde{x}(\tau),\tilde{p}(\tau)\big) := \big(x(t(\tau)),p(t(\tau))\big), \ \ 0\leqslant\tau\leqslant\int_0^T\lambda(t)\,dt$$

which are bona fide periodic solutions of (2), and which lie on $H^{-1}(c)$.

The upshot of the foregoing is the following conclusion: it suffices to prove the theorem for $H = h$ and $c = 1$. This we now do.

[3] See Clarke [12].

We define Λ to be the Fenchel conjugate of h:

$$\Lambda(v,w) = \max_{(\alpha,\beta)} \big\{ (v,w)\bullet(\alpha,\beta) - h(\alpha,\beta) \big\}.$$

Claim. Λ is C^1 and strictly convex, and, for some $\kappa > 0$, satisfies

$$\kappa\big|(v,w)\big|^2 \leqslant \Lambda(v,w) \;\; \forall\,(v,w).$$

Consider now the following isoperimetric problem: to minimize

$$\int_0^1 \Lambda\big(-y',x'\big)\,dt \;\text{ subject to } \int_0^1 y\bullet x'\,dt = 1,\; x(0)=x(1)=0,\; y(0)=y(1)=0.$$

Claim. There exists a (global) solution (x_*,y_*) of this problem. [Hint: Exer. 21.18.]

We now seek to write necessary conditions. It was pointed out in Example 17.10 that the structural hypothesis of Theorem 17.9 is satisfied. That result implies the existence of η,λ not both zero (with $\eta = 0$ or 1) such that (x_*,y_*) satisfies the integral Euler equation for the Lagrangian $\eta\,\Lambda(-y',x') + \lambda\, y\bullet x'$.

This yields constants c_1, c_2 such that

$$\eta\,\Lambda_w(-y_*',x_*') + \lambda\, y_* = c_1, \quad -\eta\,\Lambda_v(-y_*',x_*') = c_2 + \lambda\, x_* \text{ a.e.}$$

It follows that if $\eta = 0$, then y_* is constant, which is not possible in view of the isoperimetric constraint. Hence we may assume $\eta = 1$. Now if $\lambda = 0$, then $\nabla\Lambda(-y_*',x_*')$ is constant, which implies that (x_*', y_*') is constant (since Λ is strictly convex, see Exer. 4.17). This is not possible, since y_* is nonconstant and periodic.

The foregoing allows us to assert that the functions $\hat{x} = -\lambda\, x_* - c_2$, $\hat{y} = c_1 - \lambda\, y_*$ satisfy

$$(\hat{x},\hat{y}) = \nabla\Lambda\big(\hat{y}'/\lambda, -\hat{x}'/\lambda\big).$$

This is equivalent to

$$-\big(-\hat{y}',\hat{x}'\big)/\lambda = \nabla h(\hat{x},\hat{y})$$

by subdifferential inversion (see Exer. 4.27), and it follows (additional exercise) that $(\hat{x},\hat{y})$ lies on a level surface $h^{-1}(b)$ for some positive b. We now define

$$x(t) = \hat{x}\big(-t/\lambda\big)/\sqrt{b}\,, \quad y(t) = \hat{y}(-t/\lambda)/\sqrt{b}$$

if λ is negative, and otherwise we set

$$x(t) = \hat{x}\big(1-t/\lambda\big)/\sqrt{b}\,, \quad y(t) = \hat{y}(1-t/\lambda)/\sqrt{b}\,.$$

Since h is positively homogeneous of degree 2, we derive that ∇h is homogeneous of degree 1 (yet another exercise). It then follows easily that (x,y) satisfies (2) and is periodic on the interval $[0, |\,\lambda\,|]$, and that (x,y) lies on $h^{-1}(1)$. □

21.31 Exercise. (Periodic Hamiltonian trajectories 2) The previous exercise considered the question of periodic trajectories of prescribed energy. Here we study the case of prescribed *period*, in the context of Newton's equation. It is not possible to prescribe both the energy and the period (unsurprisingly).

Let $V : \mathbb{R}^n \to \mathbb{R}$ be convex and satisfy

$$V(x) \leqslant c|x|^R + d \ \ \forall x \in \mathbb{R}^n \tag{3}$$

for certain constants c, d, R with $R < 2$, together with

$$\liminf_{x \to 0} V(x)/|x|^2 > 0. \tag{4}$$

Under these hypotheses, we prove that, for any $T > 0$, there exists $x \in C^1[0,T]$ (vector-valued) with $x' \in \mathrm{AC}[0,T]$ such that

$$x''(t) \in -\partial V\big(x(t)\big) \text{ a.e.}, \ \ x(0) = x(T)\,,\ x'(0) = x'(T)\,, \tag{5}$$

and such that x has true, or minimal, period T. (We mean by this that there is no integer $k > 1$ for which $x(0) = x(T/k),\ x'(0) = x'(T/k)$.)

The proof uses (as for the prescribed energy case treated in the preceding problem) a *dual action principle*. We consider the following problem (P):

$$\text{minimize} \quad \int_0^1 \Big\{ V^*\big(-q'(t)\big) + |y'(t)|^2/2 + T q'(t) \bullet y(t) \Big\}\, dt$$

subject to zero boundary conditions on the arcs $y, q \in \mathrm{AC}[0,1]$. ($V^*$ refers to the convex conjugate of V, as usual; see §4.2.)

(a) Show by the direct method, and with the help of hypothesis (3) that this problem admits a global solution (y,q).

(b) Prove that the solution (y,q) is Lipschitz.

(c) Invoke necessary conditions to deduce that $y' \in \mathrm{AC}[0,1]$ and satisfies, for some constant α:

$$-y''(t) \in T\,\partial V\big(T\,y(t) - \alpha\big),\ \ t \in [0,1] \text{ a.e.}$$

(d) Set $x(t) = T\,y(t/T) - \alpha$, and verify that x satisfies (5).

(e) Use hypothesis (4) to show that the minimum in (P) is strictly negative.

(f) Deduce from this the minimality of the period T. □

21.32 Exercise. We study the problem of minimizing

$$\int_0^1 \Big\{ \sqrt{|x(t) - x'(t)|} + x'(t) \Big\}\, dt$$

over the arcs x on $[0,1]$ satisfying $x(0) = 0$. This is a special case of the problem treated in § 18.4 with

$$n = 1, \quad \ell \equiv 0, \quad E = \{0\} \times \mathbb{R}, \quad \Lambda(x,v) = \sqrt{|x-v|} + v.$$

Observe that Λ is continuous, but neither locally Lipschitz in x nor convex in v.

(a) Verify that Λ satisfies the growth condition of Hypothesis 18.11.

(b) Show that $x_* \equiv 0$ fails to be a strong local minimizer.

(c) Prove that x_* is a weak local minimizer. □

21.33 Exercise. Consider the Hamilton-Jacobi problem (HJ) of § 19.3. In the context of Theorem 19.10, recall that the solution $u = u_*$ is given by

$$u(\tau,\beta) = \min \ \ell\big(x(0)\big) + \int_0^\tau \Lambda\big(x(t), x'(t)\big)\, dt\,,$$

where Λ is the Lagrangian corresponding to H, and where the minimum (attained) is taken over the arcs x satisfying $x(\tau) = \beta$ (with $x(0)$ free).

(a) Let x_* be a solution of the problem defining $u(\tau,\beta)$. Prove that

$$u\big(s, x_*(s)\big) = \ell\big(x_*(0)\big) + \int_0^s \Lambda\big(x_*(t), x_*'(t)\big)\, dt\,, \ \ s \in [0,\tau].$$

Deduce that when H, x_* and u are continuously differentiable near the points in question, then

$$x_*'(s) = H_p\big(x_*(s), u_x(s, x_*(s))\big)\,.$$

We now examine the case of (HJ) in which

$$n = 1, \ H(x,p) = p^2/2 - \cos x - 1\,, \ \ell(x) = 0.$$

The corresponding Lagrangian is $\Lambda(x,v) = \cos x + v^2/2 + 1$, which is more or less that of the action integral of Example 14.6. Note that the minimum value of $\Lambda(x,v)$ is 0, attained at $x = \pi$, $v = 0$ (which, oddly perhaps, corresponds to the unstable equilibrium in which the pendulum is inverted).

We wish to prove that the solution u to (HJ) provided by Theorem 19.10 is *not* a classical one, despite the smoothness of the data. We proceed by contradiction, assuming the contrary.

(b) Prove that for τ sufficiently large, we have $u(\tau,0) < \tau$.

Let x_* be a minimizer for the problem defining $u(\tau,0)$, for a value of τ as above.

(c) Prove that $x_* \in C^1[0,\tau]$, and that there exists a constant h such that

$$x_*'(t)^2/2 - \cos x_*(t) = h \ \ \forall\, t \in [0,\tau].$$

(d) Observe that $u(\tau,\cdot)$ is an even function, so that $u_x(\tau,0) = 0$. Use this fact, together with the conclusion of (a), to deduce that $h = -1$.

(e) Obtain a contradiction to conclude the proof. □

21.34 Exercise. Let ζ be an element of the dual of $W_0^{1,p}(\Omega)$, $1 \leqslant p < \infty$, and let q be the conjugate exponent to p. Prove the existence of functions

$$f_i \in L^q(\Omega) \ (i = (0,1),\dots,n) \ \text{ such that}$$

$$\langle \zeta, u \rangle = \int_\Omega f_0(x)\,u(x)\,dx + \sum_{i=1}^n \int_\Omega f_i(x)\,D_i u(x)\,dx \ \ \forall u \in W_0^{1,p}(\Omega).$$

Show that these functions are *not* uniquely determined, however. (The dual of a Sobolev space is not particularly agreeable, alas.) □

21.35 Exercise. Let u_i be a sequence in $W^{1,p}(\Omega)$ $(1 < p < \infty)$ converging weakly to u_*, and suppose that $Du_i(x) \in C$ a.e. for each i, where C is a compact convex subset of $\mathbb{R}^n$. Prove that $Du_*(x) \in C$ a.e. □

21.36 Exercise. Let $1 \leqslant p < \infty$, and let Ω be the open unit ball in $\mathbb{R}^n$. Prove the *nonexistence* of a continuous linear operator $T : L^p(\Omega) \to L^p(\partial\Omega)$ such that

$$Tu = u|_{\partial\Omega} \ \ \forall u \in C(\overline{\Omega}) \subset L^p(\Omega).$$

Thus, the functions in $L^p(\Omega)$ do not admit natural boundary values, unlike the functions in $W^{1,p}(\Omega)$. □

21.37 Exercise. (The Neumann problem) As usual, Ω is taken to be a nonempty bounded open subset of $\mathbb{R}^n$.

(a) Show that if F satisfies the hypotheses of Theorem 20.30, then the functional $J : W^{1,r}(\Omega) \to \mathbb{R}_\infty$ defined by

$$J(u) = \int_\Omega F\big(x, u(x), Du(x)\big)\,dx$$

attains a minimum (in the absence of any prescribed boundary conditions).

(b) The absence of boundary conditions leads to additional information about the minimizer (a transversality condition). We illustrate this, in the case $n = r = 2$, $\Omega =$ the unit ball, with the following notation and Lagrangian F:

$$x = (x,y),\ z = (v,w),\ F(x,y,u,v,w) = (u^2 + v^2 + w^2)/2 - \theta(x,y)\,u,$$

where $\theta : \mathbb{R}^2 \to \mathbb{R}$ is continuous. Prove that if the minimizing function u of part (a) lies in $C^2(\overline{\Omega})$, then it satisfies

$$-\triangle u + u = \theta \ \ \forall (x,y) \in \Omega \ \text{ and } \ Du(x,y) \bullet \nu(x,y) = 0 \ \ \forall (x,y) \in \partial\Omega,$$

where $\nu(x,y)$ is the unit normal vector to Ω at (x,y). (This is known as a *Neumann boundary value problem* in the theory of partial differential equations; the notion of weak solutions can be extended to such problems.) □

21.38 Exercise. Consider the Dirichlet problem

$$\triangle u(x) = 0 \ (x \in \Omega), \quad u|_{\Gamma} = \varphi, \tag{6}$$

$$\text{where} \quad n = 2, \quad \Omega = \{(x,y) : 2x^2 + y^2 < 2\}, \quad \varphi(x,y) = 1 + 2x^2 + 3y^2.$$

Recall that in the Lipschitz context, a weak solution of (6) refers to a function u in $\operatorname{Lip}(\Omega)$ which satisfies the boundary condition as well as

$$\int_{\Omega} Du(x) \bullet D\psi(x)\, dx = 0 \ \forall \psi \in \operatorname{Lip}_0(\Omega).$$

Prove that there is a unique weak solution u of the problem (6) above, and that it satisfies $u(x,y) \geqslant 3 \ \forall (x,y) \in \Omega$. □

21.39 Exercise. We outline the proof of Theorem 20.31 (recall that Ω is bounded).

In order to apply the direct method, two celebrated theorems are required. The first of these is *Poincaré's inequality*, which asserts the existence of a constant C such that

$$\|u\|_{L^r(\Omega)} \leqslant C\|Du\|_{L^r(\Omega)} \quad \forall u \in W_0^{1,r}(\Omega).$$

The reader will note the commonality with Wirtinger's inequality of Exer. 21.14.

The second is the *Rellich-Kondrachov theorem*, which asserts that when a sequence converges weakly in $W_0^{1,r}(\Omega)$, then some subsequence converges strongly in $L^r(\Omega)$. This is akin to Exer. 6.7.

The proofs of these facts are arduous, and involve quite a bit of approximation; we admit them for the purposes of this exercise. We turn now to the proof of the theorem. It is clear that a minimizing sequence u_i exists.

(a) Show that Du_i is bounded in $L^r(\Omega)$.

(b) Use Poincaré's inequality to deduce that u_i is bounded in $L^2(\Omega)$.

(c) Prove that some subsequence (we do not relabel) is such that $u_i - \varphi$ converges weakly in $W^{1,r}(\Omega)$ to a limit $\hat{u}$. Show that $u_* := \hat{u} + \varphi$ is admissible for the basic problem.

(d) With the help of the Rellich-Kondrachov theorem, establish that some further subsequence u_i is such that u_i converges almost everywhere to u_*, and Du_i converges weakly to Du_*.

(e) Conclude by invoking the integral semicontinuity theorem 6.38. □

21.40 Exercise. Let Ω be a nonempty bounded open subset of $\mathbb{R}^n$, and let φ be a real-valued function defined on $\Gamma = \partial\Omega$. We say that φ satisfies the *lower bounded slope condition* with constant K if, given any point $\gamma \in \Gamma$, there exists an affine function of the form $y \mapsto \langle \zeta_\gamma, y-\gamma \rangle + \varphi(\gamma)$ with $|\zeta_\gamma| \leqslant K$ such that

$$\langle \zeta_\gamma, y-\gamma \rangle + \varphi(\gamma) \leqslant \varphi(y) \quad \forall y \in \Gamma.$$

This requirement can be viewed as a partial, one-sided, or lower version of the classical bounded slope condition (see p. 402). It has an alternate characterization:

Proposition. *In order for φ to satisfy the lower bounded slope condition with constant K, it is necessary and sufficient that φ be the restriction to Γ of a function $f : \mathbb{R}^n \to \mathbb{R}$ which is convex and globally Lipschitz of rank K.*

(a) Prove the sufficiency.

(b) Prove the necessity, with the help of the function f defined by

$$f(x) = \sup_{\gamma \in \Gamma} \langle \zeta_\gamma, x-\gamma \rangle + \varphi(\gamma).$$

(c) Prove that φ satisfies the bounded slope condition if and only if φ is both the restriction to Γ of a convex function $f : \mathbb{R}^n \to \mathbb{R}$, as well as the restriction to Γ of a concave function $g : \mathbb{R}^n \to \mathbb{R}$. Deduce from this that any φ satisfying the bounded slope condition must be affine when restricted to any segment contained in Γ.

As it happens, there is a version of the Hilbert-Haar theorem 20.20 in which the bounded slope condition is replaced by the *lower* bounded slope condition. Unsurprisingly, we obtain less regularity of the solution u_* than before. Nonetheless, the conclusion includes the crucial property that u_* is locally Lipschitz in Ω (but not necessarily Lipschitz on Ω); see Clarke [15] for details. □

Part IV
Optimal Control

Chapter 22

Necessary conditions

The proof of the maximum principle, given in the book of Pontryagin, Boltyanskii, Gamkrelidze and Mischenko... represents, in a sense, the culmination of the efforts of mathematicians, for considerably more than a century, to rectify the Lagrange multiplier rule.

L. C. Young
(Calculus of Variations and Optimal Control Theory)

As a child, I merely knew this; *now I can explain it.*

David Deutsch (The Fabric of Reality)

More than any other mathematical tool, it is differential equations that have been used to describe the way the physical world behaves. Systems of ordinary differential equations of the form

$$x'(t) = f\big(t, x(t)\big)$$

are routinely used today to model a wide range of phenomena, in areas as diverse as aeronautics, power generation, robotics, economic growth, and natural resources. The great success of this paradigm is due in part to the fact that it suggests a natural mechanism through which the behavior of the system can be influenced by external factors.

This is done by introducing an explicit *control variable* in the differential equation, a time-varying parameter that can be chosen (within prescribed limits) so as to attain a certain goal. This leads to the main object of our attention in this and subsequent chapters, the *controlled differential equation*

$$x'(t) = f\big(t, x(t), u(t)\big), \; u(t) \in U.$$

The couple (f, U) is referred to as the *control system*.

In the classical calculus of variations, in studying such phenomena as the shape of a soap film or the motion of a pendulum, the principal goal is descriptive: to determine the underlying governing equations, and in doing so, reveal the "natural" behavior of the system. This has typically been done with the aid of such axioms as d'Alembert's principle, or the principle of least action. In the case of the pendulum, we have seen (Example 14.6) how this leads to the differential equation

$$\theta''(t) + (g/\ell)\sin\theta(t) = 0.$$

In control theory, however, the governing equations are the *starting point* of the analysis. The system is viewed as a tool for imposing desired behavior. A canonical example in contemporary courses in control is the *inverted pendulum*, in which one seeks to design a control law for the *controlled* system

F. Clarke, *Functional Analysis, Calculus of Variations and Optimal Control*, Graduate Texts in Mathematics 264, DOI 10.1007/978-1-4471-4820-3_22,

$$\theta''(t) + (g/\ell)\sin\theta(t) = u(t)$$

having the effect that the pendulum will be driven to its unstable equilibrium $\theta = \pi$. Such undignified behavior would be considered highly inappropriate by a classical pendulum.

Thus, control theory makes a profound break, at the philosophical level, with the calculus of variations. This explains why a number of new issues of central importance arise. For example, *controllability* (the very possibility of steering the state to a given target set) becomes a key question, as well as *stabilization* and *feedback* control (in which u is a function of x). Cost functionals become more flexible: they can even be invented, in order to induce desired responses.

The issue that we turn to now is that of optimality, the one that is most closely linked to the calculus of variations. We begin the discussion of the *optimal control problem* by first introducing the reader to some standard terminology. We are given an interval $[a,b]$, the *dynamics function*

$$f : [a,b] \times \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^n,$$

as well as a subset U of $\mathbb{R}^m$, the *control set*. A *control* is a measurable function on $[a,b]$ with values in U. The *state*, or *state trajectory*, corresponding to the control u refers to a solution x of the initial-value problem

$$x'(t) = f\big(t, x(t), u(t)\big),\ x(a) = x_0, \quad t \in [a,b] \text{ a.e.},$$

where $x_0 \in \mathbb{R}^n$ is a prescribed initial condition. Thus, $x : [a,b] \to \mathbb{R}^n$ is an arc (a vector-valued function with absolutely continuous components). This differential equation linking the control u and the state x is referred to as the *state equation*. We wish to choose the control u generating the state x in such a way as to minimize a cost $J(x,u)$ defined by

$$J(x,u) = \ell\big(x(b)\big) + \int_a^b \Lambda\big(t, x(t), u(t)\big)\,dt,$$

where Λ (the *running cost*) and ℓ (the *endpoint cost*) are given functions. In so doing, we must also respect the endpoint constraint $x(b) \in E$, where E, the *target set*, is a prescribed subset of $\mathbb{R}^n$.

It is common to refer to a couple (x,u) obtained as above as a *process* of the underlying control system (f,U). Thus, a process means a pair (x,u) consisting of an arc x and a measurable function u which satisfy

$$x'(t) = f\big(t, x(t), u(t)\big),\ u(t) \in U,\ t \in [a,b] \text{ a.e.}$$

In summary then, here is the standard **optimal control problem (OC)**:

$$\left\{\begin{array}{ll}\text{Minimize} & J(x,u) = \ell\big(x(b)\big) + \int_a^b \Lambda\big(t,x(t),u(t)\big)\,dt \\ \text{subject to} & x'(t) = f\big(t,x(t),u(t)\big),\ t \in [a,b] \text{ a.e.} \\ & u(t) \in U,\ t \in [a,b] \text{ a.e.} \\ & x(a) = x_0,\ x(b) \in E.\end{array}\right. \tag{OC}$$

In this section, we shall take U to be bounded, and we posit the following standard regularity assumptions on the data:

22.1 Hypothesis. (The classical regularity hypotheses) *The function ℓ is continuously differentiable. The functions f and Λ are continuous, and admit derivatives $D_x f(t,x,u)$ and $D_x \Lambda(t,x,u)$ relative to x which are themselves continuous in all variables (t,x,u).*

These hypotheses, which are imposed globally for simplicity, imply that the cost $J(x,u)$ is well defined for any process. They do not imply that every measurable control u generates a state arc x. We could make a linear growth assumption on the dynamics function to guarantee this, but there is no need to do so at present.

Let (x_*, u_*) be a given process satisfying the constraints of (OC). We call it a *local minimizer* provided that, for some $\varepsilon > 0$, for any other process (x,u) satisfying the constraints, as well as $\|x - x_*\| \leqslant \varepsilon$, we have $J(x_*, u_*) \leqslant J(x,u)$.[1] It is also common to refer to a local minimizer (x_*, u_*) as an *optimal* (or locally optimal) process. In this terminology, u_* is an *optimal control* and x_* is an *optimal trajectory*.

We do make one exception to the classical regularity context, for the sake of efficiency. Instead of considering separately various cases in which the target set E is either a point, the whole space, a smooth manifold, or a manifold with boundary, as is often done, we simply take E, from the start, to be any closed set. In so doing, the special cases will all be appropriately subsumed.

A good understanding of optimal control begins with the study of its ancestor, the calculus of variations, which the reader is expected to be acquainted with to some extent. (If necessary, we are quite willing to pause here while the reader scans at least the first three or four chapters of Part III.)

The very terminology of optimal control is often inspired by its predecessor, an example of this phenomenon being the following.

The **Hamiltonian** function $H^\eta : [a,b] \times \mathbb{R}^n \times \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$ associated to the problem (OC) is defined by

$$H^\eta(t,x,p,u) = \langle p, f(t,x,u) \rangle - \eta\, \Lambda(t,x,u).$$

---

[1] As in the calculus of variations, the anonymous norm $\|x\|$ always refers to the relevant supremum norm, in this case $\sup_{t \in [a,b]} |x(t)|$.

The parameter η in this expression will usually be either 1 or 0, the *normal* and *abnormal* cases, respectively. When $\eta = 1$, we usually write H for H^1. The *maximized Hamiltonian* of the problem[2] is the function M^η defined by

$$M^\eta(t,x,p) = \sup_{u \in U} H^\eta(t,x,p,u).$$

These functions play a role in writing the necessary conditions that an optimal process must satisfy, a topic that we turn to now.

22.1 The maximum principle

The following seminal result dates from about 1960; it is known in the trade as the *Pontryagin maximum principle*.

22.2 Theorem. *Let the process* (x_*, u_*) *be a local minimizer for the problem* (OC) *under the classical regularity hypotheses, and where* U *is bounded. Then there exists an arc* $p : [a,b] \to \mathbb{R}^n$ *and a scalar* η *equal to* 0 *or* 1 *satisfying the* **nontriviality condition**

$$\big(\eta, p(t)\big) \neq 0 \;\; \forall\, t \in [a,b],$$

the **transversality condition**

$$-p(b) \in \eta \nabla \ell\big(x_*(b)\big) + N_E^L\big(x_*(b)\big),$$

the **adjoint equation** *for almost every* t:

$$-p'(t) = D_x H^\eta\big(t, x_*(t), p(t), u_*(t)\big),$$

as well as the **maximum condition** *for almost every* t:

$$H^\eta\big(t, x_*(t), p(t), u_*(t)\big) = M^\eta\big(t, x_*(t), p(t)\big).$$

If the problem is autonomous (that is, if f *and* Λ *do not depend on* t*), then one may add to these conclusions the* **constancy of the Hamiltonian:** *for some constant* h*, we have*

$$H^\eta\big(x_*(t), p(t), u_*(t)\big) = M^\eta\big(x_*(t), p(t)\big) = h \text{ a.e.}$$

Remarks.

(a) The maximum principle is a hybrid multiplier rule that combines conclusions of stationarity (as in Theorem 9.1) and separation (as in Theorem 9.4). Its proof, not

[2] We bow here to the usage which has come to prevail. It would be more accurate to refer to H as the *pre-Hamiltonian*. The *true* Hamiltonian of the problem is in fact M. Ah well.

an easy task, is postponed (§22.6). In the meantime, we shall strive to understand its implications, and familiarize ourselves with a few of its variants.

(b) The adjoint equation may be expressed without using H^η, as follows:

$$-p'(t) = D_x f\big(t, x_*(t), u_*(t)\big)^* p(t) - \eta\, \Lambda_x\big(t, x_*(t), u_*(t)\big),$$

where the symbol $*$ denotes the transpose (or *adjoint*) of the Jacobian matrix $D_x f$. This explains the provenance of the term "adjoint." We refer to the arc p as the *costate*; it is also known as the *adjoint arc*. The adjoint and state equations may be expressed together as follows:

$$x' = H_p^\eta(t,x,p,u), \; -p' = H_x^\eta(t,x,p,u).$$

This resembles a classical Hamiltonian system of differential equations, but with an extra control term present.

(c) We expect the reader to find the transversality condition rather familiar, in view of our earlier results in the calculus of variations. When the set E is defined by functional relations, such as equalities and inequalities, the transversality condition can be formulated equivalently in terms of multipliers; see below.

(d) In order to assert the necessary conditions of the maximum principle, it suffices, unsurprisingly, that the regularity hypotheses 22.1 hold *near* x_* and U; that is, on a set of the form

$$\big\{ (t,x,u) \,:\, t \in [a,b], \; |x - x_*(t)| < \varepsilon, \; d_U(u) < \varepsilon \big\}.$$

But we shall go well beyond this later in formulating more refined hypotheses that bear only upon the points (t,x,u) for which u lies in U. Another issue that will be addressed is that of unbounded controls.

(e) The presence of the "normality multiplier" η in the above statement of the maximum principle is familiar from the various forms of the multiplier rule seen earlier (for example, Theorem 9.1). As before, the abnormal case arises when the constraints are so restrictive (tight) that, of themselves, they identify the optimal solution regardless of the cost. (An example is given in Exer. 22.4 below.)

Note that in the abnormal case $\eta = 0$, the two components of the cost, ℓ and Λ, do not explicitly appear in the conclusions of the maximum principle. The following result asserts that this pathology does *not* happen when the final state value $x(b)$ is (locally) unconstrained.

22.3 Corollary. *In the context of* Theorem 22.2, *suppose that* $E = \mathbb{R}^n$, *or more generally, that* $x_*(b) \in \operatorname{int} E$. *Then the maximum principle holds with* $\eta = 1$.

Proof. Let us suppose that the maximum principle holds with $\eta = 0$, and obtain from this a contradiction. The transversality condition implies that $p(b) = 0$, since,

when $x_*(b) \in \operatorname{int} E$, we have $N_E^L(x_*(b)) = \{0\}$. When $\eta = 0$, the adjoint equation reduces to the following linear differential equation for p:

$$-p'(t) = D_x f\big(t, x_*(t), u_*(t)\big)^* p(t).$$

But any solution p of such a linear differential equation that vanishes at one point necessarily vanishes everywhere (by Gronwall's lemma, Theorem 6.41). Then the nontriviality condition of Theorem 22.2 is violated: contradiction. □

22.4 Exercise. Consider the following problem, in which $n = m = 2$ and $\Lambda = 0$:

$$\begin{cases} \text{Minimize} & J(x,y,u,v) = y(1) \\ \text{subject to} & \\ & (x'(t), y'(t)) = (u(t), v(t)), \ t \in [0,1] \text{ a.e.} \\ & (u(t), v(t)) \in U := \{(u,v) : u^2 + v^2 \leqslant 1\}, \ t \in [0,1] \text{ a.e.} \\ & (x(0), y(0)) = (0,0), \ x(1) = 1. \end{cases}$$

Show that (in view of the boundary conditions) the optimal control is the only feasible one: $(u(t), v(t)) = (1,0)$ a.e. Prove that the maximum principle (Theorem 22.2) holds only in abnormal form. □

We now record an alternative statement of the maximum principle, one that expresses nontriviality in a different way.

22.5 Proposition. *The conclusions of* Theorem 22.2 *are equivalent to the existence of a scalar* $\eta \geqslant 0$ *and an arc* p *that satisfy the modified nontriviality condition* $\eta + \|p\| = 1$, *together with the other conclusions.*

Proof. Suppose first that we have the assertions of Theorem 22.2. Then the scalar $r := \eta + \|p\|$ is positive, and we can redefine η and p as follows:

$$\widetilde{\eta} = \eta / r, \quad \widetilde{p} = p / r.$$

Then $(\widetilde{\eta}, \widetilde{p})$ satisfies all the other conclusions, as well as $\widetilde{\eta} + \|\widetilde{p}\| = 1$.

Conversely, let $\eta \geqslant 0$ and p satisfy $\eta + \|p\| = 1$, together with the other conclusions of the maximum principle. If $\eta > 0$, then we redefine again:

$$\widetilde{\eta} = 1, \quad \widetilde{p} = p / \eta,$$

and this yields a pair (η, p) in the sense of Theorem 22.2. If $\eta = 0$, then $\|p\| = 1$, and the only thing to check is that $p(t)$ is never 0. This follows from Gronwall's lemma, as shown in the proof of Corollary 22.3. □

Treating time as a state component. There is a well-known device by which, under certain conditions, the time variable t can be "absorbed into the dynamics" by

a reformulation. We explain this now, in the context of the problem (OC) in its full non autonomous generality.

Let us define a new (initial) state component z lying in $\mathbb{R}$, so that the new augmented state x_+ corresponds to (z,x) in $\mathbb{R}^{n+1}$. We impose the additional dynamics $z' = 1$, and we redefine the data of the optimal control problem as follows:

$$f_+(z,x,u) = \big(1, f(z,x,u)\big), \;\; \Lambda_+(z,x,u) = \Lambda(z,x,u),$$
$$U_+ = U, \;\; x_{0+} = (a,x_0), \;\; E_+ = \mathbb{R}\times E.$$

Then an augmented process (x_+, u) is admissible for the new augmented problem (where $x_+ = (z,x)$) if and only if $z(t) = t$ on $[a,b]$ and (x,u) is an admissible process for the original problem.

So we have done nothing but relabel the time t, it seems (by absorbing it in the dynamics and calling it z). However, the new augmented problem is *autonomous*. Thus, the extra conclusion asserted by the maximum principle in that case, the constancy of the Hamiltonian, is available. To assert this, however, we require that f_+ and Λ_+ be differentiable with respect to z; that is, that f and Λ be differentiable with respect to t, which was not in the original hypotheses. This explains the need for the extra hypothesis in the corollary below:

22.6 Corollary. *In addition to the hypotheses of* Theorem 22.2, *suppose that f and Λ admit derivatives with respect to t that are continuous in all variables (t,x,u). Then we may add to the conclusions the following: for every $t \in [a,b]$, we have*

$$M^{\eta}\big(t, x_*(t), p(t)\big) + \int_t^b H_t^{\eta}\big(s, x_*(s), p(s), u_*(s)\big)\,ds \;=\; M^{\eta}\big(b, x_*(b), p(b)\big),$$

where H_t^{η} refers to the partial derivative of H^{η} with respect to the t variable.

22.7 Exercise.

(a) Derive the corollary, using the reformulation device described above.

(b) Show that the conclusion coincides with constancy of the Hamiltonian as expressed in Theorem 22.2 when the problem is autonomous. □

The reformulation method is inappropriate when the t dependence of f and Λ is not differentiable; we shall see such cases later. A further type of t dependence that is not reducible in this way occurs when the control set U depends on t (even "smoothly"). If t is interpreted as a new state component z, then the constraint becomes $u \in U(z)$, in which the control set depends on the state. We examine later (§25.4) this type of *mixed constraint*, which makes the problem more complex; it is *not* covered by Theorem 22.2.

The transversality condition. In the context of Theorem 22.2, suppose that the target set E is defined by

$$E = \{ x \in S : g(x) \leqslant 0,\ h(x) = 0 \}, \tag{1}$$

where g and h are continuously differentiable functions with values in $\mathbb{R}^{k_1}$ and $\mathbb{R}^{k_2}$ respectively, and S is a closed subset of $\mathbb{R}^n$. Then the statement of the maximum principle can be formulated so as to make g and h appear explicitly in the transversality condition.

22.8 Corollary. *Let* (x_*, u_*) *be a local minimizer for the problem* (OC) *under the classical regularity hypotheses, where* U *is bounded and* E *is given by* (1). *Then there exist a costate arc* $p : [a,b] \to \mathbb{R}^n$, *a scalar* η *equal to* 0 *or* 1, *and multipliers* $\gamma \in \mathbb{R}^{k_1}$, $\lambda \in \mathbb{R}^{k_2}$ *satisfying the nontriviality, positivity, and complementary slackness conditions*

$$\big(\eta, p(t), \gamma, \lambda\big) \neq 0 \ \forall t \in [a,b],\ \ \gamma \geqslant 0,\ \ \langle \gamma, g(x_*(b)) \rangle = 0,$$

together with the **explicit transversality condition**

$$-p(b) \in \eta \nabla \ell\big(x_*(b)\big) + D_x\big\{ \langle \gamma, g \rangle + \langle \lambda, h \rangle \big\}\big(x_*(b)\big) + N_S^L\big(x_*(b)\big),$$

as well as the adjoint equation and the maximum condition.

Proof. Suppose first that the natural constraint qualification for the set E is satisfied at $x_*(b)$ (see Exer. 11.40). Then we simply invoke Theorem 22.2 directly, and the transversality condition of that theorem yields the explicit transversality condition, in view of the available characterization of N_E^L.

Otherwise, if that constraint qualification fails, then there exist γ in $\mathbb{R}^{k_1}$ and λ in $\mathbb{R}^{k_2}$, not both zero, with $\gamma \geqslant 0$ and $\langle \gamma, g(x_*(b)) \rangle = 0$, such that

$$0 \in D_x\big\{ \langle \gamma, g \rangle + \langle \lambda, h \rangle \big\}\big(x_*(b)\big) + N_S^L\big(x_*(b)\big).$$

Then we simply take $\eta = 0$ and $p \equiv 0$ to obtain all the required conclusions. □

We remark that a corresponding version of Cor. 22.8 exists when E is defined by a constraint of the form $\varphi(x) \in \Phi$; then, it is a matter of using Theorem 11.38.

The calculus of variations. An important special case of the optimal control problem (OC) is the one in which the control u coincides with the derivative x'; that is, the case in which the governing dynamics are $x' = u$, so that the function f is given by $f(t,x,u) = u$. If, in addition, we take $U = \mathbb{R}^n$, then the optimal control problem (OC) reduces to a version of the basic problem in the calculus of variations, which we have studied in depth in Part III:

$$\min \int_a^b \Lambda\big(t, x(t), x'(t)\big)\, dt \ : x \in \mathrm{AC}[a,b],\ x(a) = x_0,\ x(b) \in E.$$

We have supposed (so far) that the control set U is bounded, which is not the case here. Let us ignore this hypothesis for the moment, and assume that the maximum principle (Theorem 22.2) applies. Then we see that, in this setting, it must hold normally: with $\eta = 1$. For otherwise, the maximum condition asserts that $\langle p(t), u \rangle \leqslant \langle p(t), x_*'(t) \rangle \ \forall u$, whence $p \equiv 0$, and nontriviality is violated.

We then observe that with $\eta = 1$, the maximum condition is equivalent to the Weierstrass condition (see Theorem 15.14). If Λ is differentiable in v, it also implies $p(t) = \Lambda_v(t, x_*(t), x_*'(t))$ a.e., so that the adjoint and state equations together yield

$$\big(p'(t), p(t)\big) = \nabla_{x,v} \Lambda\big(t, x_*(t), x_*'(t)\big).$$

The reader will recognize the Euler equation in du Bois-Raymond form (see §15.1). As for the transversality condition, it already has the form that we introduced in the calculus of variations (see Theorem 18.1). Finally, in the autonomous case, the constancy of the Hamiltonian coincides with the Erdmann condition (see Prop. 14.4 or Theorem 18.13).

We may summarize this discussion by saying that, for the basic problem, the maximum principle encapsulates all the first-order necessary conditions obtained in the calculus of variations. For more complicated problems, where x' and u differ, and in which U is not the whole space, it can be shown that, in certain cases, the conclusions of the maximum principle coincide with those of the multiplier rules of the type encountered in Chapter 17. In general, however, the maximum principle goes beyond the classical multiplier rule; for example, the control set U may consist of finitely many points.

Apart from its mathematical generality, however, an equally important aspect of the maximum principle lies in its very formulation of the underlying problem, which emphasizes the *control* aspect, the control system (f, U), and the possibility of influencing a system of differential equations for certain purposes.

22.9 Example. In theory, the use of the maximum principle to characterize optimal processes is based on the following idea: we use the maximum condition to express the optimal control value $u_*(t)$ as a function of $(x_*(t), p(t))$. Then we substitute this into the state and adjoint differential equations. There results a system of $2n$ differential equations. These are accompanied by a combination of prescribed endpoint conditions and transversality conditions which, taken together, amount to $2n$ boundary conditions. In principle, then, the function (x_*, p) is determined, and therefore the optimal process as well.

We now illustrate the use of this (admittedly ideal) procedure in a simple case.

$$\begin{cases} \text{Minimize} & J(x,u) = \displaystyle\int_0^3 \big(x(t) + u(t)^2/2\big)\, dt \\ \text{subject to} & x'(t) = u(t) \in [-1,1] \text{ a.e.} \\ & x(0) = 0. \end{cases}$$

Note that, in this version of (OC), we have

$$n = m = 1, \quad U = [-1,1], \quad E = \mathbb{R}.$$

Later results on existence will imply that a (global) minimizing process (x_*, u_*) exists; let us proceed to apply the maximum principle in order to identify it.

Because the endpoint $x(3)$ is free ($E = \mathbb{R}$), we know from Cor. 22.3 that the maximum principle holds in normal form ($\eta = 1$). Thus the appropriate Hamiltonian is given by

$$H(x, p, u) = pu - x - u^2/2.$$

The maximum condition therefore concerns the maximization over the set $[-1,1]$ of the (strictly concave) function $u \mapsto pu - u^2/2$. The necessary and sufficient condition for this maximization is the stationarity condition

$$p - u_* \in N_{[-1,1]}(u_*),$$

which characterizes the unique maximizing value u_*. When $|p| \leqslant 1$, this is satisfied by taking $u_* = p \in [-1,1]$. When $p > 1$, then $p - u_*$ is necessarily positive, so that u_* must be 1 (by the stationarity); similarly, $p < -1$ implies $u_* = -1$. We summarize:

$$u_*(t) = \begin{cases} -1 & \text{if } p(t) < -1 \\ p(t) & \text{if } -1 \leqslant p(t) \leqslant +1 \\ +1 & \text{if } p(t) > +1. \end{cases}$$

The adjoint equation affirms that $-p'(t) = H_x = -1$, and transversality provides $p(3) = 0$. We deduce that $p(t) = t - 3, \; t \in [0,3]$. These facts yield the optimal control u_*:

$$u_*(t) = \begin{cases} -1 & \text{if } 0 \leqslant t < 2 \\ t-3 & \text{if } 2 \leqslant t \leqslant 3. \end{cases}$$

Of course, this determines the optimal state trajectory:

$$x_*(t) = \begin{cases} -t & \text{if } 0 \leqslant t < 2 \\ (t-3)^2/2 - 5/2 & \text{if } 2 \leqslant t \leqslant 3. \end{cases}$$

We remark that in the absence of the control constraint in the problem (that is, when we take $U = \mathbb{R}$), the solution is given by $x_*' = u_* = t - 3 \;\; \forall t$ (see Exer. 14.20), a smoother function than the optimal state x_* above. This illustrates the general rule that optimal trajectories in control are less likely to be smooth than in the calculus of variations. □

The deductive and inductive methods In seeking to solve rigorously the optimal control problem (OC), as with any optimization problem, we can argue *deductively*: we know with certainty (from existence theory) that a solution exists; we invoke

necessary conditions whose applicability is rigorously justified; we find the best (possibly the unique) candidate identified by the necessary conditions; we conclude that it's the solution. Any weak link in this chain of reasoning leads to an answer that is, in essence, no more than a guess. In the absence of some link in the chain, however, we can attempt, in various alternative ways, to prove that a given candidate is indeed a solution: this is the *inductive* approach.

It follows that if the maximum principle is being used in a deductive context, we must have a version of it that has been proved for the *same* class of controls that is used in existence theory. This motivates the need to consider measurable controls, and explains why a version of the maximum principle limited to continuous or piecewise continuous controls (for example) is of little interest. Assuming the regularity of the solution is just as logically fatal as assuming its existence.

It is true that in most of the worked-out examples that one meets, the optimal control turns out to be piecewise continuous. In Example 22.9, u_* is actually continuous. It is important to realize, however, that the argument used to identify u_* was *not* based upon any such assumption being made *prior* to the analysis. Imagine for a moment that we knew the validity of the maximum principle only for piecewise continuous controls. It may appear that little is lost in restricting attention to this class. But in fact, the effect is drastic, since we are no longer able to reason deductively. In the example above, the only assertion one could make would be the following conditional one: *if* there is an optimal control among the class of piecewise continuous controls, then it is the one we have found.

We discuss existence theory (and regularity) later on in some detail, as well as various inductive methods. Throughout the rest of this chapter, however, the focus remains on necessary conditions.

22.2 A problem affine in the control

In this section, we propose to study the following special case of the problem (OC), in which $n = m = 1$:

$$\left\{\begin{array}{ll} \text{Minimize} & J(x,u) = \displaystyle\int_0^T \{x(t)+u(t)\}\,dt \\ \text{subject to} & x'(t) = x(t)+1-x(t)\,u(t),\ t\in[0,T] \text{ a.e.} \\ & u(t)\in U = [0,3],\ t\in[0,T] \text{ a.e.} \\ & x(0) = 0. \end{array}\right.$$

Note that the terms containing the control variable are linear in u. Optimal control problems with this feature have solutions which exhibit certain characteristics, some of which we shall discover in the analysis.

A possible interpretation of the problem is that of *limiting* the growth of the state: perhaps x measures the quantity of a weed whose presence is inevitable, but that we wish to minimize. The effort cost of reducing x is measured by u, which we *also* have interest in minimizing. The problem is how to balance these contradictory goals. Note that the term xu in the dynamics reflects the fact that a given effort (rate) u is more productive (or actually, destructive) when more weeds are present; that is, when x is large.

We shall take the planning horizon T sufficiently long, to avoid merely short-term considerations; to be specific, it turns out that $T > 2\ln 2$ is relevant in this regard, as we shall see. It follows readily from existence theory (see the remark on p. 483) that an optimal process (x,u) exists; let us admit this for now. The reader will note that $x(t) > 0$ for $t > 0$, in view of the differential equation governing the state (which implies $x' > 0$ when x is near 0).

The maximum principle, in the form of Theorem 22.2, is applicable to our problem; it holds in normal form ($\eta = 1$) since the endpoint of the state is free. We proceed to show that it identifies a unique process, which, perforce, must be optimal (the deductive method at work). We write the Hamiltonian (with $\eta = 1$) and record the adjoint equation:

$$H(x,p,u) = p\,(x+1-xu)-(x+u), \;\; -p' = p-pu-1.$$

Let us define a function σ as follows: $\sigma(t) = 1+x(t)p(t)$. This is simply the coefficient of $-u$ in the Hamiltonian. Then the maximum condition implies that (almost everywhere) we have

$$u(t) = \begin{cases} 0 & \text{if } \sigma(t) > 0 \\ 3 & \text{if } \sigma(t) < 0. \end{cases} \tag{1}$$

The role of σ is that of a *switching function*. Note that σ is absolutely continuous, and that $u(t)$ is not immediately determined when $\sigma(t) = 0$. In problems in which the control u enters linearly, as in this one, the properties of the switching function usually play a central role in the analysis; the arguments employed below are typical of the ones that arise.

Using the expressions for x' and p' given by the state and adjoint differential equations, we find

$$\sigma'(t) = x(t)+p(t) \text{ a.e.} \tag{2}$$

Thus σ is continuously differentiable. From the transversality condition, we have $p(T) = 0$, and we have imposed $x(0) = 0$, whence

$$\sigma(0) = \sigma(T) = 1. \tag{3}$$

The above implies that initially, and also near T, the switching function is positive and the optimal control is zero. Could u be identically zero? (Or more precisely, equal to 0 a.e.?) If so, solving the adjoint and state equations yields

$$x(t) = e^t - 1, \;\; p(t) = 1 - e^{T-t}, \;\; \sigma(t) = e^T \{ e^{-t} - 1 + e^{t-T} \}.$$

But this expression for $\sigma(t)$ is strictly negative at $t = T/2$, because $T > 2\ln 2$ by hypothesis. This is inconsistent with $u = 0$, in light of (1). (Thus, only in the short term can it be optimal to do nothing; that is, when T is small.)

Note: This proves that there is a *first* point $\tau_1 \in (0,T)$ for which $\sigma(\tau_1) = 0$.

22.10 Proposition. *Let* $x(t) < 1$ *for some* $t \in (0,T)$*. Then we have* $\sigma(t) > 0$.

Proof. We argue by contradiction: suppose that $\sigma(t) \leqslant 0$. Then $1 + x(t)p(t) \leqslant 0$ and

$$\sigma'(t) = x(t) + p(t) \leqslant (x(t)^2 - 1)/x(t) < 0.$$

It follows that $\sigma < 0$ on an open interval $(t, t+\varepsilon)$. Since $\sigma(T) = 1$, there is a first $\tau > t$ at which $\sigma(\tau) = 0$ holds. Clearly, we must have $\sigma'(\tau) \geqslant 0$.

In (t,τ), then, we have $\sigma < 0$, whence $u = 3$ and $x' = -2x + 1$. Since $x(t) < 1$ (by assumption), we deduce from this differential equation that

$$x(s) \leqslant \max(1/2, x(t)) < 1, \;\; s \in [t, \tau].$$

However, the conditions $\sigma'(\tau) \geqslant 0$ and $\sigma(\tau) = 0$ together imply $x(\tau) \geqslant 1$, which yields the desired contradiction. □

Note: It follows that for the time τ_1 defined previously, we have $x(\tau_1) \geqslant 1$.

22.11 Proposition. *Let* $x(t) > 1$ *for some* $t \in (0,T)$*. Then for* $s \in (t, T]$ *we have* $\sigma(s) > 0$, *and consequently* $u(s) = 0$ a.e.

Proof. Observe that when $x > 1/2$ and $\sigma < 0$, then (locally) $u = 3$ and $x' < 0$ a.e. It follows that for some $\bar{t} < t$, we must have

$$x(\bar{t}) > 1 \text{ and } \sigma(\bar{t}) \geqslant 0.$$

Now suppose that the conclusion of the proposition is false. We shall derive a contradiction, considering first the case in which $\sigma(\bar{t}) > 0$.

Then there exists a first $\tau > \bar{t}$ for which $\sigma(\tau) = 0$; we must have $\sigma'(\tau) \leqslant 0$. In $(\bar{t}, \tau)$ we have $\sigma > 0$, whence $u = 0$ and x is increasing. Thus we have $x(\tau) > 1$. Combined with $\sigma(\tau) = 0$, this gives $\sigma'(\tau) > 0$, the required contradiction.

Consider finally the case $\sigma(\bar{t}) = 0$. This equality, combined with $x(\bar{t}) > 1$, implies $\sigma'(\bar{t}) > 0$. But then, for a positive ε sufficiently small, we have $x(\bar{t} + \varepsilon) > 1$ and $\sigma(\bar{t} + \varepsilon) > 0$. The argument of the first case above now applies (for $\bar{t}$ replaced by $\bar{t} + \varepsilon$), and yields the required contradiction. □

Note: The proposition implies that once $x(t) > 1$, this persists thereafter, with corresponding control value $u = 0$ a.e. We deduce from this that $x(\tau_1) = 1$, as follows.

We already know that $x(\tau_1) \geqslant 1$, and if we had $x(\tau_1) > 1$ then we would have $u = 0$ a.e. both before and after τ_1, a possibility that has been ruled out.

The same reasoning shows that there cannot be points $\tau_1 + \varepsilon$ with $\varepsilon > 0$ arbitrarily small such that $x(\tau_1 + \varepsilon) > 1$. We now examine the opposite possibility.

22.12 Proposition. *There is no $t > \tau_1$ for which $x(t) < 1$.*

Proof. Suppose to the contrary that there is such a value of t. Then there is a greatest value $\tau < t$ for which $x(\tau) = 1$. There must be a set S of positive measure contained in the interval (τ, t) such that $x(s) < 1$ and $x'(s) < 0$ for $s \in S$. But according to Prop. 22.10, we have $x' > 0$ a.e. when $x < 1$: contradiction. □

It follows from the above that there cannot be points $\tau_1 + \varepsilon$ with $\varepsilon > 0$ arbitrarily small such that $x(\tau_1 + \varepsilon) < 1$.

Conclusion. As a consequence of the above, we deduce that there is a maximal interval of the form $[\tau_1, \tau_2]$ (with $\tau_2 > \tau_1$) on which x is identically 1; we must have $x' = 0$ on that interval, which implies $u = 2$ a.e. there. On the interval $[\tau_1, \tau_2]$, the switching function σ must vanish; $[\tau_1, \tau_2]$ is referred to as a *singular interval*. We necessarily have $\tau_2 < T$, since $\sigma(T) = 1$. Subsequently, we have $x > 1$ and $u = 0$ a.e. on the interval $(\tau_2, T]$.

We now have a clear qualitative picture of the optimal process (see Fig. 22.1). The optimal control turns out to be piecewise constant, with two switches.

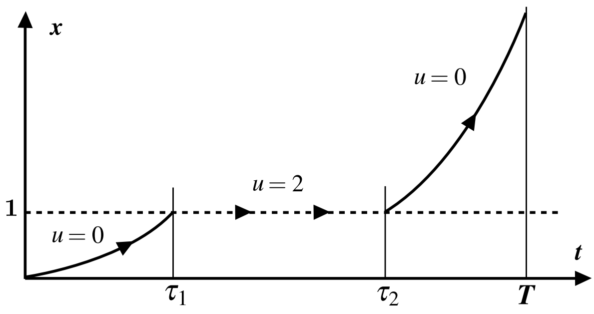


Fig. 22.1
The turnpike solution

The privileged value $x = 1$ corresponds to what is called a *turnpike*, which reflects long-term optimality. The optimal strategy consists of heading to the turnpike as fast as possible (here, with $u = 0$) and then staying there, until the terminal time is approached, at which point short-term considerations take over, and no further effort is applied.

There remains to calculate τ_1 and τ_2; this information is also implicit in the necessary conditions. As for τ_1, it is simply the first time at which $x = 1$, using $u = 0$. We find, from the differential equation for the state, $\tau_1 = \ln 2$.

Let us now calculate τ_2. Note that in (τ_1, τ_2), we have $p = -1$ (as follows from $\sigma = 0$ and $x = 1$). On $[\tau_2, T]$, however, we have

$$p' = 1 - p, \;\; p(T) = 0 \implies p(t) = 1 - e^{T-t}, \;\; t \in [\tau_1, \tau_2].$$

The point τ_2 is then seen to be the value of t for which $1 - e^{T-t} = -1$: we find that $\tau_2 = T - \ln 2$. The reader will observe that when T is large, the optimal state trajectory is at the turnpike value most of the time.

22.3 Problems with variable time

An important feature of certain optimal control problems is that the underlying interval is itself a choice variable. We refer to a problem in which the interval $[a, b]$ is not prescribed, but has some possibility of varying, as a *variable-time problem.* A canonical example of this type is the *minimal-time problem.* It consists of finding the process (x, u) on an interval $[0, \tau]$ (where τ is not specified beforehand) which attains $x(\tau) = 0$, and where τ is the *least* time for which this is possible. (Thus we are seeking the quickest trajectory to the origin.)

We consider now the following variable-time optimal control problem (VT), in which the terminal time τ may be free to vary:

$$\begin{cases} \text{Minimize} & J(\tau, x, u) = \ell\big(\tau, x(\tau)\big) + \int_0^\tau \Lambda\big(x(t), u(t)\big)\,dt \\ \text{subject to} & \tau \geqslant 0 \\ & x'(t) = f\big(x(t), u(t)\big), \;\; t \in [0, \tau] \text{ a.e.} \\ & u(t) \in U, \;\; t \in [0, \tau] \text{ a.e.} \\ & x(0) = x_0, \;\; \big(\tau, x(\tau)\big) \in S. \end{cases} \qquad \textbf{(VT)}$$

We have underlined the dependence on the *horizon* τ (as it is sometimes called) by writing $J(\tau, x, u)$; the endpoint cost function ℓ now depends on the horizon τ as well. Observe that problem (VT) reduces to a fixed-time problem of the type considered earlier if S is of the form $\{T\} \times E$: then the horizon is explicitly prescribed. At the other extreme (no restriction on terminal time) is the minimal-time problem, which corresponds to taking $S = \mathbb{R}_+ \times \{0\}$, $\ell \equiv 0$, $\Lambda \equiv 1$. (It is equivalent to take $\ell \equiv \tau$ and $\Lambda \equiv 0$.) We stress that f and Λ are *autonomous* here: independent of t. In this case, there is no loss of generality in taking the initial point of the underlying interval (formerly known as a) to be 0, as we have done.

Basic hypotheses. In the theorem below, ℓ, f and Λ are assumed to satisfy the same classical regularity hypotheses 22.1 as before. We also continue to assume that U is bounded. Thus, the cost integral is well defined for any process satisfying the problem constraints.

Before stating necessary conditions for the problem (an appropriately modified version of the maximum principle), we need to extend the concept of local minimum to variable-time problems, which we now do. To measure the "closeness" of an arc x_1 defined on an interval $[0,\tau_1]$ to that of an arc x_2 defined on an interval $[0,\tau_2]$, we simply extend each of the arcs to $[0,\infty)$ by constancy; thus, for example, we set $x_1(t) = x_1(\tau_1)\ \forall\, t \geqslant \tau_1$. This convention induces a meaning for the expression $\|x_1 - x_2\|$, which here refers to $\max_{t\,\geqslant\,0} |x_1(t) - x_2(t)|$.

Let the process (x_*, u_*), defined on the interval $[0,\tau_*]\,(\tau_* > 0)$, satisfy the constraints of (VT). We say it is a **local minimizer** if, for some $\varepsilon > 0$, for all processes (x,u) on an interval $[0,\tau]$ satisfying the constraints of (VT) as well as

$$|\tau - \tau_*| \leqslant \varepsilon \ \text{ and } \ \|x - x_*\| \leqslant \varepsilon,$$

we have $J(\tau_*, x_*, u_*) \leqslant J(\tau, x, u)$.

In the following variable-time maximum principle, the Hamiltonian H^η and the maximized Hamiltonian M^η are defined exactly as before:

$$H^\eta(x,p,u) = \langle p, f(x,u)\rangle - \eta\,\Lambda(x,u), \quad M^\eta(x,p) = \sup_{u\,\in\,U} H^\eta(x,p,u).$$

22.13 Theorem. *Let the process* (x_*, u_*)*, defined on the interval* $[0,\tau_*]$ $(\tau_* > 0)$ *be a local minimizer for the problem* (VT) *under the hypotheses above. Then there exists an arc* $p : [0,\tau_*] \to \mathbb{R}^n$ *and a scalar* η *equal to* 0 *or* 1 *satisfying the* **nontriviality condition**

$$\big(\eta, p(t)\big) \neq 0 \ \ \forall\, t \in [0,\tau_*],$$

the **adjoint equation** *for almost every* $t \in [0,\tau_*]$:

$$-p'(t) = D_x H^\eta(x_*(t), p(t), u_*(t)),$$

as well as the **maximum condition:** *for almost every* $t \in [0,\tau_*]$,

$$H^\eta\big(x_*(t), p(t), u_*(t)\big) = M^\eta\big(x_*(t), p(t)\big),$$

and such that, for some constant h, *we have* **constancy of the Hamiltonian:**

$$H^\eta\big(x_*(t), p(t), u_*(t)\big) = M^\eta\big(x_*(t), p(t)\big) = h \text{ a.e.},$$

and the **transversality condition**

$$\big(h, -p(\tau_*)\big) \in \eta\nabla\ell\big(\tau_*, x_*(\tau_*)\big) + N_S^L\big(\tau_*, x_*(\tau_*)\big).$$

Note that the constant value h of the Hamiltonian now figures in the transversality condition. When S is of the form $\{T\}\times E$ (as was the case earlier), then the transversality condition yields no information on h, and reduces to precisely the transversality condition of Theorem 22.2. At the other extreme, when S is of the form $\mathbb{R}_+ \times E$ and ℓ does not depend on τ (as in the minimal-time problem), it yields the precise value $h = 0$. Theorem 22.13 will turn out to be a special case of a more general result (see Theorem 22.22).

22.14 Example. (Soft landing) The following minimal-time problem is an obligatory feature of every introduction to optimal control; it is variously known as the robot car, double integrator, or soft landing problem. It is the simplest interesting case of the general minimal-time problem.

Adhering to tradition, then, we consider the dynamics

$$x''(t) = u(t) \in [-1,+1],$$

the goal being to find the control u that steers the initial state /velocity pair (x_0, v_0) to rest at the origin (that is, $x = x' = 0$) in least time. To express the problem in the standard formulation, which involves only first-order differential equations, we must introduce another explicit state variable, the velocity y, so that the second-order equation above takes the form of a first-order system:

$$\begin{bmatrix} x'(t) \\ y'(t) \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} x(t) \\ y(t) \end{bmatrix} + \begin{bmatrix} 0 \\ 1 \end{bmatrix} u(t), \;\; u(t) \in [-1,+1].$$

Thus we are dealing with a linear system, with $n = 2$ and $m = 1$. In the notational context of problem (VT) (p. 449), we may take

$$\ell(\tau, x) = \tau, \;\; \Lambda = 0, \;\; S = \mathbb{R}_+ \times \{(0,0)\}.$$

We now apply Theorem 22.13. The Hamiltonian is $H^{\eta}(x,y,p,q,u) = py + qu$, so that the adjoint system is given by

$$p'(t) = 0, \;\; -q'(t) = p(t).$$

It follows that $p(t)$ is a constant p_0, and that q is an affine function. The constancy of the Hamiltonian H yields $p_0\, y(t) + |q(t)| = h$ for some constant h. Transversality implies $h = \eta$. It follows that if q is identically zero, then so is p. But then we would have $h = \eta = 0$, which violates nontriviality. We conclude that q is not identically zero, and thus (being affine) changes sign at most once in $[0, \tau]$.

The maximum condition then implies that the optimal control u is *bang-bang*; that is, $u = \pm 1$ (depending on the sign of q). More explicitly, the optimal control is equal to 1 almost everywhere up to a certain point, then -1 a.e. thereafter, or else the reverse. In other words, u is piecewise constant, with values in $\{-1,+1\}$, and exhibits at most one change in sign.

The trajectories (x,y) for the constant control value $u=1$ lie on parabolas of the form $2x=y^2+c$, since we have $2x'-2yy'=0$; the movement is upward since $y'=1$. Similarly, the trajectories for $u=-1$ correspond to (leftward opening) parabolas $2x=-y^2+c$, with downward motion (see Fig. 22.2).

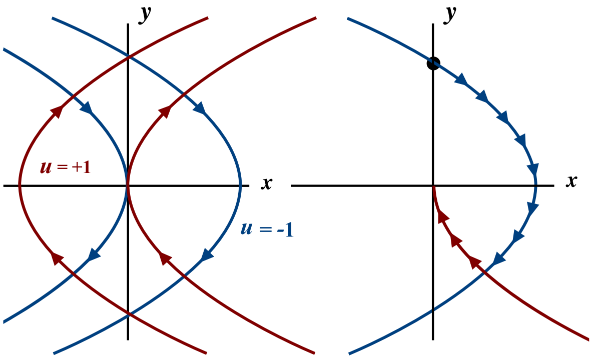


Fig. 22.2 Sample trajectories for $u \equiv +1$ and for $u \equiv -1$, and an optimal trajectory starting from the positive y-axis.

There is a unique strategy that combines two such displacements so as to reach the origin. Beginning, for example, at a point on the positive y-axis (as indicated on Fig. 22.2), we must follow a downward parabola ($u=-1$) until its intersection with the unique upward parabola passing through $(0,0)$; then we proceed to follow that one to the origin ($u=+1$).

The overall conclusion may be usefully expressed in terms of the *switching curve* Σ

$$\Sigma = \{(-y^2/2,y) : y \geqslant 0\} \bigcup \{(y^2/2,y) : y \leqslant 0\},$$

defined as the union of the two parabolic halves that are used to attain the origin. Then the (supposedly) time-optimal strategy u_* can be summarized in *synthesized* or *feedback* terms as follows (see Fig. 22.3):

$$u_* = \begin{cases} +1 & \text{if } (x,y) \text{ lies to the left of } \Sigma \\ -1 & \text{if } (x,y) \text{ lies to the right of } \Sigma. \end{cases}$$

It is further understood[3] here that on the upper branch of Σ, one uses $u=-1$, and on the lower branch, $u=+1$.

[3] While it is clear enough how this summarizes optimal behavior, the exact meaning of this feedback law as a dynamical system is somewhat unclear. The subject of "discontinuous feedback" addresses such issues.

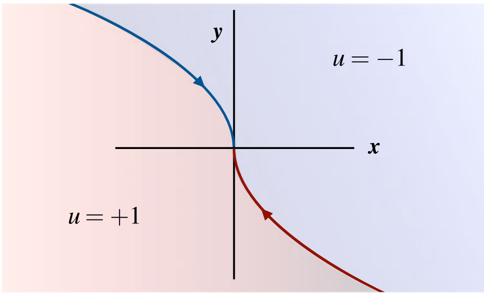


Fig. 22.3 The switching curve and the time-optimal feedback synthesis

We conclude that if time-optimal trajectories exist (for every initial condition), then they are described as above. It so happens that an existence theorem to be seen later does apply to this example (namely Theorem 23.13, but let's take our word for it for now). Given this, the deductive method assures us that we have, in fact, identified the optimal trajectories.[4]

At this point, it is a matter of routine calculation to derive an explicit formula for the corresponding optimal time as a function of the initial point (x,y). Letting this optimal time be denoted $T(x,y)$, we find:

$$T(x,y) = \begin{cases} -y+\sqrt{2y^2-4x} & \text{if } (x,y) \text{ lies to the left of } \Sigma \\ +y+\sqrt{2y^2+4x} & \text{if } (x,y) \text{ lies to the right of } \Sigma. \end{cases}$$

It is of interest (and not difficult) to show that this *minimal-time function* T is continuous, and that T is smooth on the open set which is the complement of the switching curve Σ. However, T is nondifferentiable, and indeed, fails to be locally Lipschitz, at points on Σ. □

22.15 Exercise. (Very soft landing) Consider the system

$$x'''(t) = u(t) \in [-1,1],$$

and the problem of steering the state to rest at 0 in minimal time T, in the sense that the position x, the velocity x', and the acceleration x'' must all equal 0 at T. Show that an optimal control is bang-bang with at most two switches. □

[4] We should mention that minimal-time problems with linear dynamics, of which the soft landing problem is but one example, can be studied on a systematic basis using time reversal and a technique called "backing out of the origin." See Lee and Markus [30].

22.4 Unbounded control sets

We return to the standard optimal control problem (OC) (p. 437), under the same classical regularity hypotheses as before, but now without assuming that the control set U is bounded. There is a slight complication in this extended setting: when u is unbounded, the integral term in the cost (that is, the integral of the function $t \mapsto \Lambda(t, x(t), u(t))$) may not be well defined. We deal with this by assigning an appropriate meaning to the word "admissible."

An **admissible process** for (OC) is one that satisfies the constraints of the problem, *and* for which the cost functional $J(x,u)$ is well defined. (This second requirement was automatically satisfied previously, when U was taken to be bounded.) This convention, which interprets admissibility quite permissively, is always used from now on. It allows us to discuss optimal control problems without imposing additional structural hypotheses to guarantee that certain integrals exist, while being consistent with such hypotheses if we wish to use them.

The meaning of local minimum is adapted to this new nomenclature in the natural way: We say that the process (x_*, u_*) is a local minimizer for (OC) provided that it is admissible, and that, for some $\varepsilon > 0$, for any other admissible process (x,u) satisfying $\|x - x_*\| \leqslant \varepsilon$, we have $J(x,u) \geqslant J(x_*, u_*)$.

When the control set U is unbounded, the maximum principle (Theorem 22.2) *fails* unless a compensating assumption is made. We shall use one that is reminiscent of the Tonelli-Morrey growth condition in the calculus of variations.

22.16 Hypothesis. *There exist $\varepsilon > 0$, a constant c, and a summable function d such that, for almost every $t \in [a,b]$, we have*

$$|x - x_*(t)| \leqslant \varepsilon \implies \big|D_x(f,\Lambda)\big(t,x,u_*(t)\big)\big| \leqslant c\,\big|(f,\Lambda)\big(t,x,u_*(t)\big)\big| + d(t)\,.$$

(Recall that the norm $|M|$ of a $k \times \ell$ matrix M is defined to be the Euclidean norm of its entries viewed as an element of $\mathbb{R}^{k\ell}$.)

Let us observe that this condition concerns only the control u_*, and is localized around x_*. Many systems satisfy *globally* a structural hypothesis of the type

$$|D_x f(t,x,u)| + |D_x \Lambda(t,x,u)| \leqslant c\{|f(t,x,u)| + |\Lambda(t,x,u)|\} + d(t)\,,$$

or, at least, satisfy this whenever x is restricted to a bounded set and u lies in U. Evidently, when this is the case, Hypothesis 22.16 cannot help but hold. Another instance of note concerns *bounded* controls: under classical regularity, Hypothesis 22.16 automatically holds if u_* happens to be bounded: take $c = 0$ and $d(t)$ a sufficiently large constant.[5] Thus, the next result subsumes Theorem 22.2.

---

[5] This is the compensating assumption used by Pontryagin and his collaborators.

22.17 Theorem. *Let* (x_*, u_*) *be a local minimizer for* (OC)*, under the classical regularity hypotheses, where* U *is not necessarily bounded. If* Hypothesis 22.16 *holds, then the conclusions of* Theorem 22.2 *are satisfied.*

As with the other variants of the classical maximum principle, we shall derive this theorem later from the extended maximum principle of §22.6. In the meantime, we illustrate its use by obtaining as an immediate consequence the isoperimetric multiplier rule stated in Chapter 17. The proof illustrates the well-known (but useful) technique of "absorbing an integral into the dynamics."

22.18 Corollary. Theorem 17.9 *holds.*

Proof. We augment the state x by an additional coordinate y, and introduce the augmented control system

$$[x'(t), y'(t)] = f_+\big(t, x(t), y(t), u(t)\big) := \big[u(t), \psi\big(t, x(t), u(t)\big)\big], \; u(t) \in U := \mathbb{R}^n.$$

Consider the corresponding augmented problem of type (OC), in which we set

$$\Lambda_+(t,x,y,u) = \Lambda(t,x,u), \;\; \big(x(a), y(a)\big) = (A,0),$$
$$E = \big\{(B,0)\big\}, \;\; \ell_+\big(x(b), y(b)\big) = 0.$$

(The reader will observe that this is where the isoperimetric constraint is absorbed by the augmented dynamics: it corresponds to $y(a) = 0$, $y(b) = 0$.) Now define

$$y_*(t) = \int_a^t \psi(s, x_*(s), x_*'(s))\, ds, \;\; u_*(t) = x_*'(t).$$

It is a simple bookkeeping exercise to verify that the augmented process (x_*, y_*, u_*) provides a local minimum for the augmented problem. Furthermore, Hypothesis 22.16 (in the context of this augmented problem) is equivalent to the structural growth hypothesis of Theorem 17.9, so that Theorem 22.17 may be applied. It is then a routine matter to show that its conclusions translate into the stated ones for the original isoperimetric one. □

22.19 Example. (and Exercise) We examine now a type of problem in which unbounded control sets are a natural feature. The *linear-quadratic regulator* refers to the following problem that arises in engineering applications:

$$\left\{\begin{array}{lll} \text{Minimize} & J(x,u) = \displaystyle\int_0^T \tfrac{1}{2}\langle Qu(t), u(t)\rangle\, dt & \\ \text{subject to} & x'(t) = Ax(t) + Bu(t), \;\; t \in [0,T] \text{ a.e.} & \textbf{(LQR)} \\ & u(t) \in U := \mathbb{R}^m, \;\; t \in [0,T] \text{ a.e.} & \\ & x(0) = x_0, \;\; x(T) = x_T. & \end{array}\right.$$

We assume that the $m \times m$ matrix Q is positive definite and symmetric. The $n \times n$ matrix A and the $n \times m$ matrix B are also given, together with the horizon $T > 0$ and the prescribed endpoints $x_0, x_T \in \mathbb{R}^n$. The usual interpretation of the problem is one of seeking least-energy transfer of the state between prescribed values.

Under these hypotheses, standard existence results assure us that an optimal process (x_*, u_*) exists (see Theorem 23.11). Since the constraints are linear and the cost is strictly convex, the solution (x_*, u_*) of (LQR) is unique. (This will also result from the analysis below.) But we do not know (yet) that u_* is bounded. It is an advantage of Theorem 22.17 that it does not require this assumption.

- Prove that Hypothesis 22.16 is satisfied.

In view of this, and because the classical regularity hypotheses hold, it follows that Theorem 22.17 applies, so that the conclusions of the maximum principle are available to us. The Hamiltonian is given by

$$H^\eta(t,x,p,u,\eta) = \langle p, Ax + Bu \rangle - \frac{\eta}{2} \langle Qu, u \rangle.$$

The adjoint equation of the maximum principle is $-p'(t) = A^* p(t)$, a differential equation whose solution is of the form $p(t) = e^{-A^* t} p_0$ (matrix exponential) for some $p_0 \in \mathbb{R}^n$. The maximum condition asserts that, for almost every t, the point $u_*(t)$ maximizes over $u \in \mathbb{R}^m$ the function

$$u \mapsto \langle B^* p(t), u \rangle - \frac{\eta}{2} \langle Qu, u \rangle.$$

If $\eta = 0$, then the function $t \mapsto B^* p(t) = B^* e^{-A^* t} p_0$ must be identically zero. This possibility will now be excluded by a hypothesis bearing upon the *controllability matrix*, which refers to the $n \times mn$ matrix C defined as follows:

$$C = \begin{bmatrix} B & AB & \dots & A^{n-1}B \end{bmatrix}.$$

- Prove that if $B^* e^{-A^* t} p_0$ is identically zero on $[0, T]$, then $C^* p_0 = 0$.

Let us now postulate that C has maximal rank (as is customary). Then the discussion above implies that the maximum principle cannot hold abnormally. For then we would have $p_0 = 0$, whence $p \equiv 0$, violating the nontriviality condition. Thus we have $\eta = 1$.

- With $\eta = 1$, show that the maximum condition implies

$$u_*(t) = Q^{-1} B^* p(t) = Q^{-1} B^* e^{-A^* t} p_0 \text{ a.e.}$$

This characterization evidently implies that u_* is bounded (so this is a *conclusion* rather than an assumption). Substituting in the state equation, we find

$$x'(t) = Ax(t) + BQ^{-1} B^* e^{-A^* t} p_0.$$

The variation of parameters formula then leads to

$$x(t) = e^{At}x_0 + \int_0^t e^{A(t-s)}BQ^{-1}B^*e^{-A^*s}p_0\,ds\,.$$

- Prove that the following matrix is positive definite:

$$\int_0^T e^{-As}BQ^{-1}B^*e^{-A^*s}\,ds\,.$$

  Deduce from this that the equation $x(T) = x_T$ determines p_0.

Thus, the optimal u_* and x_* are completely determined. □

22.5 A hybrid maximum principle

The uses of control theory being so diverse, there exist many variants of the standard optimal control problem. The one we treat now is a representative of a certain class of problems said to be *hybrid* because they involve several control systems.

We consider a *two-stage* control problem, in which two different standard (autonomous) control systems

$$\big(x, f(x,u), U\big) \text{ and } \big(y, g(y,v), V\big)$$

are linked over a specified planning period $[0,T]$, in the sense that the second system takes over from the first at a certain *switching time* τ; thus we have

$$\begin{aligned} x'(t) &= f\big(x(t),u(t)\big),\;\; u(t) \in U,\;\; 0 \leqslant t < \tau \text{ a.e.}\\ y'(t) &= g\big(y(t),v(t)\big),\;\; v(t) \in V,\;\;\; \tau < t \leqslant T \text{ a.e.}\end{aligned}$$

An initial condition is imposed on the first state, and a terminal condition on the second:

$$x(0) \in E_1,\;\; y(T) \in E_2\,.$$

In addition, at the switching time τ, certain linking conditions must hold:

$$\big(\tau, x(\tau), y(\tau)\big) \in S,$$

where S is a given set. Note that the switching time τ may vary; it is a choice variable. It is not assumed that the dimensions of x and y, or of u and v, are the same. Thus E_1 and E_2, for example, are subsets of generally different Euclidean spaces $\mathbb{R}^{n_1}$ and $\mathbb{R}^{n_2}$.

The minimization involves both endpoint and running cost components of a familiar type, and depends as well on the value of $(\tau, x(\tau), y(\tau))$. To summarize, here is the

hybrid optimal control problem (HC) in compact notation:

$$\left\{\begin{array}{ll}\text{Minimize} & J = \ell_1\big(x(0)\big)+\ell_2\big(y(T)\big)+\ell_0\big(\tau,x(\tau),y(\tau)\big) \\ & \qquad + \int_0^\tau \Lambda_1\big(x(t),u(t)\big)\,dt + \int_\tau^T \Lambda_2\big(y(t),v(t)\big)\,dt \\ \text{subject to} & \tau \in [0,T] \\ & x'(t) = f\big(x(t),u(t)\big),\ \ u(t)\in U,\ \ t\in[0,\tau] \text{ a.e.} \\ & y'(t) = g\big(y(t),v(t)\big),\ \ v(t)\in V,\ \ t\in[\tau,T] \text{ a.e.} \\ & x(0)\in E_1,\ \ \big(\tau,x(\tau),y(\tau)\big)\in S,\ \ y(T)\in E_2 .\end{array}\right. \tag{\textbf{HC}}$$

As in the case of the variable-time problem (VT) considered in §22.3, we need to specify the sense in which a solution to this problem is local. Let $(x_1,u_1),(y_1,v_1)$ and $(x_2,u_2),(y_2,v_2)$ be two admissible hybrid processes, with switching times τ_1 and τ_2 respectively. As we did before, we consider that x_1 is defined beyond τ_1 (to the right) by constancy; similarly for x_2 beyond τ_2. As for y_1 and y_2, they are extended to the left by constancy.

Then it makes sense to write such terms as $\|x_1-x_2\|$, taken with respect to $[0,T]$. Accordingly, a *local minimizer* for (HC) refers to an admissible hybrid process $(x_*,u_*),(y_*,v_*)$, with switching time τ_*, such that, for some $\varepsilon>0$, for any admissible hybrid process $(x,u),(y,v)$, with switching time τ, that satisfies

$$\|x-x_*\|<\varepsilon,\quad \|y-y_*\|<\varepsilon,\quad |\tau-\tau_*|<\varepsilon,$$

we have $J(x_*,y_*,\tau_*)\leqslant J(\tau,x,y)$.

In a break with prior developments, we shall now forsake the classical regularity hypotheses. Instead, we suppose that all the functions involved (that is, f,g,Λ_1,Λ_2, ℓ_0,ℓ_1,ℓ_2) are locally Lipschitz. (Of course, continuous differentiability would imply this.) We also assume that the control sets U and V are bounded, and that the sets E_1,E_2, and S are closed.

We define the usual Hamiltonians for each of the individual processes involved:

$$H_1^\eta(x,p,u) = \langle p, f(x,u)\rangle - \eta\,\Lambda_1(x,u),\ \ H_2^\eta(y,q,v) = \langle q, g(y,v)\rangle - \eta\,\Lambda_2(y,v),$$

and similarly for the maximized Hamiltonians M_1^η and M_2^η, obtained by maximizing over U and V respectively.

22.20 Theorem. *Let* (x_*,u_*) *and* (y_*,v_*)*, with switching time* $\tau_*\in(0,T)$*, be a local minimizer for the problem* (HC). *Then there exist arcs* p *on* $[0,\tau_*]$ *and* q *on* $[\tau_*,T]$ *and a scalar* η *equal to* 0 *or* 1 *satisfying the* **nontriviality condition**

$$\big(\eta,\,p(\tau_*),\,q(\tau_*)\big)\neq 0,$$

the **transversality conditions**

$$p(0) \in \eta\, \partial_L \ell_1\big(x_*(0)\big) + N^L_{E_1}\big(x_*(0)\big)\,,$$
$$-q(T) \in \eta\, \partial_L \ell_2\big(y_*(T)\big) + N^L_{E_2}\big(y_*(T)\big)\,,$$

the **adjoint inclusions**

$$-p'(t) \in \partial_C H_1^\eta\big(\bullet, p(t), u_*(t)\big)\big(x_*(t)\big)\,,\; t \in [0,\tau_*] \text{ a.e.}$$
$$-q'(t) \in \partial_C H_2^\eta\big(\bullet, q(t), v_*(t)\big)\big(y_*(t)\big)\,,\; t \in [\tau_*,T] \text{ a.e.},$$

as well as the **maximum conditions**: *for almost every* t,

$$t \in [0,\tau_*] \implies H_1^\eta\big(x_*(t), p(t), u_*(t)\big) = M_1^\eta\big(x_*(t), p(t), u_*(t)\big)$$
$$t \in [\tau_*,T] \implies H_2^\eta\big(y_*(t), q(t), v_*(t)\big) = M_2^\eta\big(y_*(t), q(t), v_*(t)\big).$$

In addition, there exist constants h_1, h_2 *such that the* **constancy conditions** *hold*:

$$H_1^\eta\big(x_*(t), p(t), u_*(t)\big) = M_1^\eta\big(x_*(t), p(t)\big) = h_1,\; t \in [0,\tau_*] \text{ a.e.}$$
$$H_2^\eta\big(y_*(t), q(t), v_*(t)\big) = M_2^\eta\big(y_*(t), q(t)\big) = h_2,\; t \in [\tau_*,T] \text{ a.e.},$$

together with the **switching condition**:

$$\big(h_1 - h_2, -p(\tau_*), q(\tau_*)\big) \in \eta\, \partial_L \ell_0\big(\tau_*, x_*(\tau_*), y_*(\tau_*)\big) + N^L_S\big(\tau_*, x_*(\tau_*), y_*(\tau_*)\big).$$

The statement of the theorem is a fearsome sight to behold. But the reader may discern that it merely speaks of two maximum principles, with a coherent link between them provided by the switching condition. The theorem is also the first (but not the last) to admit *nonsmooth* dependence on x. This is reflected by the generalized gradient ∂_C that appears in the adjoint inclusions.[6] The first of these (for example) refers to the generalized gradient of the function $x \mapsto H_1^\eta(x, p(t), u_*(t))$, evaluated at $x_*(t)$. Note that the inclusion reduces to the usual adjoint equation if the data happen to be continuously differentiable in x.

The theorem will be proved in §22.6, with the help of an extended maximum principle that we shall meet later on.

22.21 Exercise. Suppose that the second state continues the first, in the sense that $S = \big\{ (\tau, x, y) : x = y \big\}$ and $\ell_0 = 0$. (Thus, the switching time merely changes the dynamics.) Prove that, in Theorem 22.20, q continues p; that is, $p(\tau_*) = q(\tau_*)$, and $h_1 = h_2$. □

As a special case of Theorem 22.20, we obtain Theorem 22.13 for the variable-time problem (VT) considered earlier (p. 449), and extended to nonsmooth data:

[6] The nonsmooth maximum principle fails with ∂_L in the adjoint inclusion (see Exer. 22.29).

22.22 Theorem. *Let* (x_*, u_*) *on the interval* $[0, \tau_*]$ *be a local minimizer for the variable-time problem* (VT)*, where* f *and* ℓ *are locally Lipschitz,* U *is bounded, and* S *is closed. Then there exists an arc* $p : [0, \tau_*] \to \mathbb{R}^n$ *and a scalar* η *equal to* 0 *or* 1 *satisfying the* **nontriviality condition**

$$\big(\eta, p(t)\big) \neq 0 \ \ \forall t \in [0, \tau_*],$$

the **adjoint inclusion** *for almost every* $t \in [0, \tau_*]$:

$$-p'(t) \in \partial_C H^{\eta}\big(\bullet, p(t), u_*(t)\big)\big(x_*(t)\big),$$

as well as the **maximum condition:** *for almost every* $t \in [0, \tau_*]$,

$$H^{\eta}\big(x_*(t), p(t), u_*(t)\big) = M^{\eta}\big(x_*(t), p(t)\big),$$

and such that, for some constant h*, we have* **constancy of the Hamiltonian:**

$$H^{\eta}\big(x_*(t), p(t), u_*(t)\big) = M^{\eta}\big(x_*(t), p(t)\big) = h \text{ a.e.},$$

and the **transversality condition**

$$\big(h, -p(\tau_*)\big) \in \eta\, \partial_L \ell\big(\tau_*, x_*(\tau_*)\big) + N_S^L\big(\tau_*, x_*(\tau_*)\big).$$

Proof. Pick $T > \tau_*$. Then (VT) is equivalent to the special case of the hybrid problem (HC) in which the second stage is trivial and has no effect on the cost. To make this precise, let us take

$$g = 0,\ \Lambda_2 = 0,\ \ell_2 = 0,\ E_2 = \mathbb{R}^n,\ \Lambda_1 = \Lambda,\ E_1 = \{x_0\},\ \ell_1 = 0.$$

We also take $\ell_0 = \ell$ and $S = S$ (so to speak), by simply ignoring the redundant y variable. Then it is a routine matter to check that the conclusions of Theorem 22.20 reduce to the ones stated in the theorem (with $h_1 = h$), since q and h_2 are easily seen to be zero. □

A variable-horizon hybrid problem. The hybrid optimal control problem (HC) above can be generalized in several ways.[7] For example, the number of different stages (or control systems) can be greater than two, so that several switches are involved; the data can depend on t; the starting time of one stage can differ from the end time of the preceding one. In any case, control theory is constantly generating models with new features, so there is no hope of stating an ultimate, all-encompassing version of the necessary conditions. At best, we can be prepared to anticipate what they may assert, and be equipped to prove them.

Let us consider one such extension, where the only new element, relative to (HC), is that the horizon T is a choice variable, rather than being prescribed. Thus the

[7] See Clarke-Vinter [19] for details.

endpoint constraint for the second stage, $y(T) \in E_2$, is replaced by one of the form $(T, y(T)) \in E_2$, and ℓ_2 becomes a function of $(T, y(T))$. The freedom now given to T is reflected, as we would expect, in a modified transversality condition. We shall omit the proof of the following, which is very close to that of Theorem 22.20.

22.23 Theorem. *Let (x_*, u_*) and (y_*, v_*), with horizon $T_* > 0$ and switching time $\tau_* \in (0, T_*)$, be locally optimal for the problem above. Then the conclusions of* Theorem 22.20 *hold with $T = T_*$, and with the second transversality condition replaced by the following*:

$$\big(h_2, -q(T_*)\big) \in \eta\, \partial_L \ell_2\big(T_*, y_*(T_*)\big) + N^L_{E_2}\big(T_*, y_*(T_*)\big).$$

22.24 Example. (Robot arm) Let us consider the problem of controlling a robot arm so that it transfers an object from one location to another, in minimal time. If the mass of the object is not negligible, then different dynamics apply to the system before and after the object is transfered. This gives rise to a hybrid optimal control problem.

A simple model with one spatial variable x considers the case in which the arm is initially at rest at the origin. It must be guided to $x = L > 0$, where the mass changes (from m_1 to m_2), and then back to rest at the origin. The time τ at which L is reached corresponds to the switch in the underlying dynamics. The governing equation is $m_1 x'' = u$ up to time τ, and $m_2 x'' = u$ afterwards.

Notation: The spatial state component $x(t)$ is continuous throughout. It is natural, therefore, to use the same symbol x for the spatial state before and after the switch time (rather than using, say, x_1 and y_1). We shall introduce a new state variable v to represent the velocity of x and, again, retain the notation v throughout for the velocity. But one must bear in mind that (unlike x), v may have a discontinuity at τ.

The robot arm problem (RA) may then be summarized as follows:

$$\left\{\begin{array}{ll} \text{Minimize} & J = T = \displaystyle\int_0^\tau 1\,dt + \int_\tau^T 1\,dt \\ \text{subject to} & \tau \geqslant 0,\ T \geqslant 0,\ \tau \leqslant T \\ & (x', v')(t) = \begin{cases} \big(v, u/m_1\big) \text{ a.e.} & \text{if } t \in [0, \tau) \\ \big(v, u/m_2\big) \text{ a.e.} & \text{if } t \in (\tau, T] \end{cases} \\ & u(t) \in [-1, +1] \text{ a.e.} \\ & x(\tau-) = x(\tau+) = L,\ \ v(\tau+) = K v(\tau-) \\ & \big(x(0), v(0), x(T), v(T)\big) = (0,0,0,0). \end{array}\right. \qquad \textbf{(RA)}$$

Note that both T and τ are choice variables in this problem, which is therefore a variable-horizon two-stage hybrid optimal control problem of the type considered

above, one in which the set E_2 is given by $\mathbb{R}\times\{(0,0)\}$. The positive parameters m_1, m_2, L are given, as well as the positive parameter K. Existence theory can be applied to prove that there is an optimal hybrid control.

Before commencing the analysis, let us explain the meaning of K and the condition $v(\tau+) = K\,v(\tau-)$ by identifying three special cases of interest:

- *Drop-off* ($m_1 > m_2$). The object is carried to $x = L$ and dropped. Here, we take $K = 1$.
- *Hard pickup* ($m_1 < m_2$). The object, which is initially at rest, is picked up at $x = L$. Now we impose $m_1 x'(\tau-) = m_2 x'(\tau+)$ (conservation of momentum). In terms of v, this amounts to $v(\tau+) = K\,v(\tau-)$, where $K = m_1/m_2 < 1$.
- *Soft pickup* ($m_1 < m_2$). In this case the speed of the object is matched to that of the arm for the pickup. We have $x'(\tau-) = x'(\tau+)$, so that $K = 1$.

The necessary conditions of Theorem 22.20 provide costate arcs that we shall denote by (p,q), both after and before the switching time τ; thus, as for the state component v, the costate is considered to have a possible discontinuity at τ. The adjoint equation and maximum condition yield

$$-p' = 0, \;\; -q' = p, \;\; u = \begin{cases} 1 & \text{if } q > 0 \\ -1 & \text{if } q < 0. \end{cases}$$

The transversality condition of Theorem 22.23 implies that $h_2 = 0$. Then the constancy and switching conditions (together with the remaining transversality condition, which yields $h_1 = 0$) lead to

$$p(t)\,v(t) - \eta = \begin{cases} -|q(t)|/m_1 & \text{if } t < \tau \\ -|q(t)|/m_2 & \text{if } t > \tau \end{cases}, \quad q(\tau-) = K\,q(\tau+).$$

The nontriviality condition asserts

$$\big(\eta, p(\tau-), q(\tau-), p(\tau+), q(\tau+)\big) \neq 0.$$

Claim 1. The arc $q(t)$ is not identically zero on $(0,\tau)$.

For suppose that this is the case. Then, by the above, we have $p = 0$ on $(0,\tau)$, whence $\eta = 0$. We also derive $q(\tau+) = 0$. Then the nontriviality condition implies $p(\tau+) \neq 0$. It also follows that $q(t) = p(\tau+)(\tau - t)$ for $t \in (\tau, T\,]$. This contradicts $p(\tau+)\,v(T) = -|\,q(T)|/m_2$ (since $v(T) = 0$).

By an analogous argument we obtain:

Claim 2. $q(t)$ is not identically zero on (τ, T).

It follows from these two claims that q consists of two affine parts, with isolated zeros, and with $q(\tau-)$ and $q(\tau+)$ having the same sign (or both being zero). This implies that u is piecewise constant, with values ± 1 and at most two switches; thus, v' is piecewise constant.

Since $L > 0$, there must be a subinterval I of $(0,\tau)$ in which we have $v(t) > 0$ and $v'(t) > 0$. Thus $q(t) > 0$ and $u = 1$ in I. In order to steer (x,v) to $(0,0)$ from such points, the control u will have to switch values twice at points beyond I: once to $u = -1$ (in order to make $v < 0$, so that x can begin returning to 0), and then later back to $u = 1$ (so that v can be made zero at T). To allow these two switches to happen, we must have

$$q(0) > 0\,,\;\; q(\tau-) < 0\,,\;\; \text{and}\;\; q(\tau+) < 0\,,\;\; q(T) > 0\,.$$

Then necessarily $p(\tau-) > 0$ and $p(\tau+) < 0$. Invoking the equality

$$\begin{aligned} \eta \;=\; p(\tau-)\,v(\tau-)+|\,q(\tau-)|/m_1 \;&=\; p(\tau+)\,v(\tau+)+|\,q(\tau+)|/m_2 \\ &=\; p(\tau+)\,K\,v(\tau-)+|\,q(\tau-)|/(K m_2)\,, \end{aligned}$$

we arrive at

$$v(\tau-)\big\{K\,p(\tau+)-p(\tau-)\big\} \;=\; |\,q(\tau-)|(K\,m_2-m_1)/(K\,m_1 m_2)\,.$$

Note that the coefficient of $v(\tau-)$ is strictly negative.

We may draw some qualitative conclusions from the last relation. In the special case termed *Drop-off*, we have $m_1 > m_2$ and $K = 1$. It follows that $v(\tau) > 0$: the drop occurs while the arm is on its outward journey, while its velocity is positive.

A *Soft pickup* ($m_1 < m_2$, $K = 1$), on the other hand, is made on the return journey, while the arm has negative velocity.

In the case of *Hard pickup* ($m_1 < m_2$, $K = m_1/m_2$), we have $v = 0$ at τ: the arm has zero velocity at pickup, so that impact does not occur even though it is permitted. We remark that the analysis can be continued in all cases in order to obtain explicit formulas for τ, T, and $v(\tau-)$ (see [19]). □

22.6 The extended maximum principle

In this section, we present and discuss a generalization of the classical maximum principle that extends it in several ways. The theorem will be used later to derive all the variants that we have encountered so far in this chapter. It bears upon the following problem (EC):

$$\begin{cases} \text{Minimize} & J(x,u) = \ell\big(x(a),x(b)\big) + \int_a^b \Lambda\big(t,x(t),u(t)\big)\,dt \\ \text{subject to} & x'(t) = f\big(t,x(t),u(t)\big),\ t \in [a,b] \text{ a.e.} \\ & u(t) \in U(t),\ t \in [a,b] \text{ a.e.} \\ & \big(x(a),x(b)\big) \in E. \end{cases} \tag{EC}$$

The underlying interval $[a,b]$ of the problem is fixed, and the **basic hypotheses** on the problem data are the following:

- The functions $f(t,x,u)$ and $\Lambda(t,x,u)$ are LB measurable in (t,u) for each x;
- The multifunction $U(\cdot)$ is LB measurable;
- The set $E \subset \mathbb{R}^n \times \mathbb{R}^n$ is closed, and ℓ is locally Lipschitz.

The LB measurability of $U(\cdot)$ means that its graph, the set

$$\operatorname{gr} U = \big\{(t,u) \in [a,b] \times \mathbb{R}^m : u \in U(t)\big\},$$

is LB measurable; that is, measurable with respect to the σ-algebra generated by Lebesgue subsets of $[a,b]$ and Borel subsets of $\mathbb{R}^m$. The reader will recall that we discussed LB measurability at length in §6.3, where it was proved, for example, that the hypothesis on f is satisfied if, for each x, the function $(t,u) \mapsto f(t,x,u)$ happens to be measurable in t and continuous in u.

A process (x,u) is termed *admissible* for (EC) if all the constraints are satisfied and $J(x,u)$ is well defined and finite; (x_*,u_*) is said to be a local minimizer provided that, for some $\varepsilon > 0$, for every admissible process (x,u) satisfying $\|x - x_*\| \leqslant \varepsilon$, we have $J(x,u) \geqslant J(x_*,u_*)$.

The main hypothesis of the theorem concerns local Lipschitz behavior of f and Λ with respect to the state:

22.25 Hypothesis. *There exists an* LB *measurable function* $k(t,u) : \operatorname{gr} U \to \mathbb{R}$ *such that, for almost every* t *in* $[a,b]$, *we have*

$$x,y \in B(x_*(t),\varepsilon),\ u \in U(t) \implies$$
$$|f(t,x,u) - f(t,y,u)| + |\Lambda(t,x,u) - \Lambda(t,y,u)| \leqslant k(t,u)|x-y|,$$

and such that $t \mapsto k(t,u_*(t))$ *is summable.*

As usual, we define the Hamiltonians H^η and M^η by

$$H^\eta(t,x,p,u) = \langle p, f(t,x,u)\rangle - \eta\,\Lambda(t,x,u),$$
$$M^\eta(t,x,p) = \sup_{u \in U(t)} H^\eta(t,x,p,u).$$

22.26 Theorem. (Clarke) *Let* (x_*, u_*) *be a local minimizer for* (EC)*, where* Hypothesis 22.25 *holds. Then there exist an arc* $p : [a,b] \to \mathbb{R}^n$ *together with a scalar* η *equal to* 0 *or* 1 *satisfying the* **nontriviality condition**

$$\big(\eta, p(t)\big) \neq 0 \;\; \forall t \in [a,b], \tag{N}$$

the **transversality condition**

$$\big(p(a), -p(b)\big) \in \eta\, \partial_L \ell\big(x_*(a), x_*(b)\big) + N_E^L\big(x_*(a), x_*(b)\big), \tag{T}$$

the **adjoint inclusion** *for almost every* t:

$$-p'(t) \in \partial_C H^\eta\big(t, \bullet, p(t), u_*(t)\big)(x_*(t)), \tag{A}$$

as well as the **maximum condition** *for almost every* t:

$$H^\eta\big(t, x_*(t), p(t), u_*(t)\big) = M^\eta\big(t, x_*(t), p(t)\big). \tag{M}$$

If the problem is autonomous (*that is, if* f, U, *and* Λ *do not depend on* t)*, then one may add to these conclusions the* **constancy of the Hamiltonian**: *for some constant* h*, we have*

$$H^\eta\big(x_*(t), p(t), u_*(t)\big) = M^\eta\big(x_*(t), p(t)\big) = h \text{ a.e.}$$

At this point, we expect the conclusions of this *extended maximum principle* to have a familiar appearance to the reader. Let us dwell for a moment on the ways in which the word "extended" is justified, relative to the classical maximum principle (Theorem 22.2).

- The nature of the boundary cost, as well as the boundary condition, is more general than before.
- u_* is not assumed to be bounded.
- The control set U may not be bounded, and is allowed to depend on t.
- The behavior of f and Λ with respect to t and u is measurable, not necessarily continuous. Only the values of these functions on the control set itself are used in formulating the main hypothesis.
- The functions involved are not assumed to be differentiable; the limiting subdifferential ∂_L and the generalized gradient ∂_C are called upon to express the conclusions, which reduce to the earlier ones if the data happen to be smooth.

It turns out that all these features are of interest, for various reasons, some of which we illustrate in this section, and others in the exercises. The proof of the extended maximum principle will be given later, as an outgrowth of results on differential inclusions, in §25.2.

22.27 Exercise. Suppose that in the context of Theorem 22.26, we have

$$x_*(b) \in \operatorname{int}\{x_1 \in \mathbb{R}^n : \exists\, x_0 \in \mathbb{R}^n \text{ such that } (x_0, x_1) \in E\}.$$

Prove that the theorem necessarily holds with $\eta = 1$. □

22.28 Example. The following simple problem features a nondifferentiable cost integrand:

$$\begin{cases} \text{Minimize} & J(x,u) = c\,x(1) + \int_0^1 |x(t)|\,dt \\ \text{subject to} & x'(t) = u(t),\ t \in [0,1] \text{ a.e.} \\ & u(t) \in [-1,+1],\ t \in [0,1] \text{ a.e.} \\ & x(0) = \alpha. \end{cases}$$

It follows from existence theory (see Theorem 23.11) that there is an optimal process (x,u) for the problem. It is clear that the hypotheses of Theorem 22.26 are satisfied. In applying it, we may take $\eta = 1$ (by the exercise above), since the endpoint $x(1)$ is unrestricted.

Now let p be the costate provided by the theorem. Then we find, from the constancy of the Hamiltonian:

$$H(x,p,u) = pu - |x| \implies \max_{u \in U} H(x,p,u) = |p| - |x| \implies |p(t)| - |x(t)| = h,$$

for some constant h. From the adjoint inclusion and the maximum condition, we deduce

$$x'(t) = u(t) = \begin{cases} -1 & \text{if } p(t) < 0 \\ +1 & \text{if } p(t) > 0 \end{cases} \qquad p'(t) = \begin{cases} -1 & \text{if } x(t) < 0 \\ +1 & \text{if } x(t) > 0. \end{cases}$$

The movement of the arc $(x(t), p(t))$ in the x-p phase plane is governed by these dynamics, and it is restricted to a level set of the form $|p| - |x| = h$; see Fig. 22.4. If $x(t)$ vanishes on an interval $[t_1, t_2]$, then necessarily $u(t) = 0$ a.e. on the interval, so that $p(t)$ also vanishes on $[t_1, t_2]$ (since $|u| = 1$ when $p \neq 0$). It follows that the level set contains the origin, whence $h = 0$. Otherwise, for $h \neq 0$, there can be no pauses in the motion.

The transversality condition provides two boundary conditions: $x(0) = \alpha$ and $p(1) = -c$. This information identifies the unique solution of the problem. □

22.29 Exercise. Show that when $\alpha < 0$ and $c < 0$, then the solution x exhibits an interval in which $x(t) = 0$ if and only if $|\alpha| + |c| < 1$. Find the optimal state trajectory when we take $\alpha = c = -1/3$. Observe that in this case, the necessary

conditions of Theorem 22.26 would fail to hold if ∂_C were replaced by ∂_L in the adjoint inclusion. □

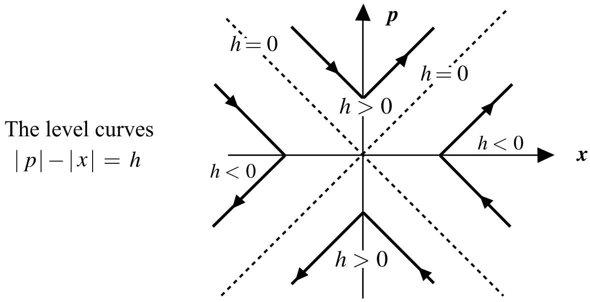


Fig. 22.4 Some typical level curves

Constant control sets. When the control set U is constant, the hypothesis of LB measurability in Theorem 22.26 forces U to be a Borel set. While this is not so very restrictive, it can be dispensed with when f and Λ are continuous in u, as is the case, notably, when the classical hypotheses hold.

We proceed to record this variant of the theorem, in a somewhat more general case that postulates the following structure:

22.30 Hypothesis. *There exists a finite or countable family* $\{(t_i, t_{i+1})\}$ *of disjoint open intervals with*

$$\operatorname{cl} \bigcup_{i \geqslant 1} (t_i, t_{i+1}) = [a, b]$$

such that on each interval (t_i, t_{i+1})*, the control set* $U(t)$ *is a constant set* U_i *(with no measurability restriction). Furthermore, for almost every* $t \in (t_i, t_{i+1})$*, for every* $x \in B(x_*(t), \varepsilon)$*, the functions* $u \mapsto f(t, x, u)$ *and* $u \mapsto \Lambda(t, x, u)$ *are continuous on the set* U_i*.*

22.31 Corollary. *Let the data satisfy all the hypotheses of* Theorem 22.26*, except that the* LB *measurability of* $U(\cdot)$ *is replaced by the hypothesis above. Then the conclusions of the theorem continue to hold.*

We remark that when the family $\{(t_i, t_{i+1})\}$ contains a single element (so that there is no actual partition of the underlying interval $[a, b]$) and when the data are autonomous, the conclusions once again include the constancy of the Hamiltonian. The proof of the corollary is postponed to a convenient moment in §25.2.

Two derivations. We proceed now to kill four birds with two stones. We shall use the extended maximum principle to derive Theorem 22.17, which (as we have pointed out) subsumes the classical maximum principle (Theorem 22.2); then, we

shall derive Theorem 22.20 for hybrid systems, which includes the variable-time problem as a special case (Theorem 22.22). There will remain, of course, to prove the extended maximum principle, as well as its corollary above. This will be done in Chapter 25.

Derivation of Theorem 22.17. For each $u \in U$, the quantity

$$k(t,u) = \max\left\{ |D_x\big(f,\Lambda\big)(t,x,u)| : x \in B(x_*(t),\varepsilon)\right\}$$

provides a Lipschitz constant for the function $x \mapsto (f,\Lambda)(t,x,u)$ on $B(x_*(t),\varepsilon)$, as required by Hypothesis 22.25. It follows that k is LB measurable, since the supremum defining k is equivalent to a countable one, by the usual argument involving a dense set. In order to justify applying Cor. 22.31 (with no actual partition of the underlying interval), which gives directly the desired conclusions, we need only verify that $t \mapsto k(t,u_*(t))$ is summable.

Fix t for which Hypothesis 22.16 holds, and any unit vector v, and define

$$g(r) = \big|(f,\Lambda)\big(t,x_*(t)+rv,u_*(t)\big)\big|, \;\; 0 \leqslant r \leqslant \varepsilon.$$

Then g is Lipschitz and satisfies

$$|g'(r)| \leqslant c\,g(r)+d(t), \;\; r \in [0,\varepsilon] \text{ a.e.},$$

as a consequence of Hypothesis 22.16. It follows from Gronwall's lemma that

$$g(r) \leqslant \alpha\big|(f,\Lambda)\big(t,x_*(t),u_*(t)\big)\big|+\beta, \;\; 0 \leqslant r \leqslant \varepsilon$$

for certain constants α and β independent of t and v. This implies in turn

$$\big|(f,\Lambda)\big(t,x,u_*(t)\big)\big| \leqslant \alpha\big|(f,\Lambda)\big(t,x_*(t),u_*(t)\big)\big|+\beta \;\; \forall x \in B(x_*(t),\varepsilon).$$

Substituting into Hypothesis 22.16, we discover

$$\big|D_x\big(f,\Lambda\big)\big(t,x,u_*(t)\big)\big| \leqslant c\,\alpha\big|(f,\Lambda)\big(t,x_*(t),u_*(t)\big)\big|+c\,\beta+d(t)$$

for all $x \in B(x_*(t),\varepsilon)$. Since the function of t on the right side is summable, we deduce that $t \mapsto k(t,u_*(t))$ is summable, as required. This completes the proof.

Derivation of the hybrid maximum principle. We now use the extended maximum principle in order to derive Theorem 22.20; the proof will demonstrate the utility of admitting discontinuous dependence with respect to t.

The plan of this fairly involved (but in no way profound) proof is based on defining a new, augmented, non hybrid problem for which the control (u_*,v_*) is optimal. Then the extended maximum principle is invoked for this problem, and the resulting necessary conditions are reinterpreted in original terms in order to obtain the stated conclusions.

The augmented problem has four states (x,y,α,β) and three controls (u,v,w), where the new state components α,β and the new control w are all real-valued. The data of the problem, whose underlying interval is $[0,T]$, are given by:

$$f_+\big(t,x,y,\alpha,\beta,u,v,w\big) = \begin{cases} \big(w\,f(x,u),0,w,w\big) & \text{if } t\in[0,\tau_*) \\ \big(0,w\,g(y,v),0,w\big) & \text{if } t\in(\tau_*,T] \end{cases}$$

$$U_+ = U\times V\times[1-\delta,1+\delta]$$

$$\ell\big(x_0,y_0,\alpha_0,\beta_0,x_1,y_1,\alpha_1,\beta_1\big) = \ell_1(x_0)+\ell_2(y_1)+\ell_0\big(\alpha_1,x_1,y_0\big)$$

$$\Lambda\big(t,x,y,\alpha,\beta,u,v,w\big) = \begin{cases} w\Lambda_1(x,u) & \text{if } t\in[0,\tau_*) \\ w\Lambda_2(y,v) & \text{if } t\in(\tau_*,T]. \end{cases}$$

Here, δ is any sufficiently small number in $(0,1/2)$ to be prescribed later. The corresponding augmented cost is denoted by J_+. We shall write the endpoint constraints in the following order for convenience:

$$\big(x(0),y(T),\alpha(0),\beta(0),\beta(T),\alpha(T),x(T),y(0)\big) \in E$$
$$:= E_1\times E_2\times\{0\}\times\{0\}\times\{T\}\times S.$$

We may consider that u_* and v_* are defined on $[0,T]$ (any extension with values in U and V respectively will do), and we extend x_* to $[0,T]$ by the constant value $x_*(\tau_*)$; similarly, we set $y_*(t)$ equal to $y_*(\tau_*)$ for $t\in[0,\tau_*]$. We further define

$$w_*(t) = 1,\ \ \beta_*(t) = t,\ \ t\in[0,T],\quad \alpha_*(t) = \begin{cases} t & \text{if } t\in[0,\tau_*) \\ \tau_* & \text{if } t\in(\tau_*,T]. \end{cases}$$

Note then that the control (u_*,v_*,w_*) and corresponding state $(x_*,y_*,\alpha_*,\beta_*)$ on $[0,T]$ are admissible for the augmented problem.

Claim. The control (u_*,v_*,w_*) and state $(x_*,y_*,\alpha_*,\beta_*)$ provide a local minimum for the augmented problem, if δ is sufficiently small.

We proceed to prove this by contradiction. Note that the cost $J(*)$ of the original process equals the cost $J_+(*)$ of the augmented version.

Suppose that (x,y,α,β) and (u,v,w) constitute an admissible process that is better (that is, assigns a lower value to J_+) for the augmented problem, where

$$\|x-x_*\| < \varepsilon/2,\ \ \|y-y_*\| < \varepsilon/2,\ \ \|\alpha-\alpha_*\| < \varepsilon/2.$$

We shall rescale time by means of the bi-Lipschitz (Lipschitz with Lipschitz inverse) transformation r from $[0,T]$ to $[0,T]$ given by

$$r(t) = \int_0^t w(s)\,ds.$$

(Note that $r(T) = T$ as a result of the constraint $\beta(T) = T$.) Set $\widetilde{\tau} = \alpha(\tau_*)$. When restricted, $r(\cdot)$ is a bi-Lipschitz transformation from $[0,\tau_*]$ to $[0,\widetilde{\tau}]$. We proceed to define an arc $\widetilde{x}$ and a measurable function $\widetilde{u}$ on $[0,\widetilde{\tau}]$ by

$$\widetilde{x}(r) = x\big(t(r)\big), \;\; \widetilde{u}(r) = u\big(t(r)\big),$$

where $t(r)$ refers to the inverse function of $r(t)$. Then, for almost every $r \in [0,\widetilde{\tau}]$, we have

$$(d/dr)\,\widetilde{x}(r) = x'\big(t(r)\big)/w\big(t(r)\big) = f\big(x(t(r)),u(t(r))\big) = f\big(\widetilde{x}(r),\widetilde{u}(r)\big),$$

so that the dynamics of the original problem are satisfied on $[0,\widetilde{\tau}]$.

Similarly, we define, for $r \in [\widetilde{\tau},T]$:

$$\widetilde{y}(r) = y\big(t(r)\big), \;\; \widetilde{v}(r) = v\big(t(r)\big).$$

Then $(\widetilde{x},\widetilde{y})$ and $(\widetilde{u},\widetilde{v})$ define a hybrid process, with switching time $\widetilde{\tau}$, which is admissible for the original problem. We find (with the help of the change of variables formula for integrals) that its cost J is that of J_+ for the augmented process (x,y,α,β), (u,v,w), a cost which is less than $J(*)$ by assumption. This contradicts the optimality of (x_*,y_*,u_*,v_*), provided of course that $(\widetilde{x},\widetilde{y})$ and $\widetilde{\tau}$ are sufficiently close to (x_*,y_*) and τ_* respectively. Let us see how to arrange this.

For any $r \in [0,\widetilde{\tau}]$, we have

$$|\widetilde{x}(r) - x_*(r)| = |x(t(r)) - x_*(r)| \leqslant |x(t(r)) - x_*(t(r))| + |x_*(t(r)) - x_*(r)|.$$

The first term on the right is bounded above by $\varepsilon/2$ by assumption, and so is the second, provided that δ has been chosen small enough, in a way that depends only on x_*. (This is an evident consequence of the uniform continuity of x_*.) Similar assertions hold for $\widetilde{y}$ and $\widetilde{\tau}$. The claim is proved.

It is easy to see that Cor. 22.31 applies to the augmented problem (with no actual partition of the underlying interval). We deduce from this result the existence of a costate arc, which we choose to write as (p,q,p_3,p_4), together with a scalar η equal to 0 or 1, for which

$$(\eta, p(t), q(t), p_3(t), p_4(t)) \neq 0 \;\; \forall t \in [0,T],$$

and satisfying the remaining conclusions.

The augmented Hamiltonian $H_+^{\eta}(t,x,y,\alpha,\beta,p,q,p_3,p_4,u,v,w)$ is given by

$$H_+^{\eta} = \begin{cases} w\,\langle p, f(x,u)\rangle + w(p_3+p_4) - \eta\, w\Lambda_1(x,u) & \text{if } t < \tau_* \\ w\,\langle q, g(y,v)\rangle + w\,p_4 - \eta\, w\Lambda_2(y,v) & \text{if } t > \tau_*. \end{cases}$$

It follows from the augmented adjoint equation that p_3, p_4 are constant, and that p and q satisfy the desired adjoint equations.

The maximum conditions follow as well, by fixing $w = 1$ in the augmented maximum condition. Furthermore, by setting $u = u_*$, $v = v_*$ and letting w vary, we derive from the augmented maximum condition that, almost everywhere on $[0, \tau_*]$, the maximum over $w \in [1-\delta, 1+\delta]$ of the function

$$w \mapsto w\{\langle p(t), f(x_*(t), u_*(t))\rangle - \eta\Lambda_1(x_*(t), u_*(t)) + p_3 + p_4\}$$

occurs at $w = 1$. Necessarily, then, we have

$$H_1^\eta\big(x_*(t), p(t), u_*(t)\big) = h_1 := -(p_3 + p_4),\ t \in [0, \tau_*] \text{ a.e.},$$

whence the first constancy condition. The second one follows similarly, with h_2 revealed to be $-p_4$.

Consider now the nontriviality condition of the augmented problem, which, substituting for p_3, p_4 from above, may now be seen as asserting that

$$\big(\eta, p(t), q(t), h_2 - h_1, -h_2\big) \neq 0\ \forall t \in [0, T].$$

We observe that (as a result of the augmented adjoint equation) p is constant on $[\tau_*, T]$ and q is constant on $[0, \tau_*]$. Now suppose that $\eta = 0$. Then it is a consequence of Gronwall's lemma (and the adjoint inclusion) that p is either never zero or else identically zero, and similarly for q. When both p and q are zero, the constancy of the Hamiltonians implies that h_1 and h_2 are zero. It follows from these facts that the nontriviality assertion above is equivalent to that of Theorem 22.20.

The remaining conclusions to be verified, namely the transversality conditions and the switching condition, are direct consequences of the transversality condition for the augmented problem. The proof is complete.

Chapter 23

Existence and regularity

Up to now, our study of optimal control has focused on the issue of necessary conditions: the maximum principle and its variants. We turn now to the attendant issues of existence and regularity. Before doing so, however, we digress somewhat in order to discuss an entirely new consideration, one that is associated with the agreeable term *relaxation*.

23.1 Relaxed trajectories

23.1 Example. Consider the following special case of the standard optimal control problem (OC) (p. 437), in which the dimension of the state (x,y,z) is 3, the control u is one-dimensional, and the underlying interval is $[0,1]$:

$$\begin{cases} \text{Minimize} & J(x,y,z,u) = x(1) \\ \text{subject to} & x'(t) = \cos u(t), \\ & y'(t) = \sin u(t), \\ & z'(t) = y(t)^2, \\ & u(t) \in U = [-\pi/2, \pi/2], \\ & x(0) = y(0) = z(0) = z(1) = 0. \end{cases}$$

The solution of the problem is evident, if one reasons as follows. Since $z' = y^2 \geqslant 0$, then, in view of the boundary conditions on z, the only admissible state trajectory (x,y,z) has $z \equiv 0$, whence $y \equiv 0$, which implies $u(t) = 0$ a.e., so that $x' = 1$ a.e., which leads to an optimal cost value $J = 1$.

Let us imagine now the possibility of violating the endpoint constraint, to any arbitrarily small positive tolerance ε, by allowing $z(1) \leqslant \varepsilon$. Then y may be taken to be a sawtooth function having zero endpoints that satisfies $|y(t)| \leqslant \varepsilon$ on $[0,1]$ as well as $y' = \pm 1$ a.e. This implies $u(t) = \pm\pi/2$ a.e., which yields in turn $x' = 0$ a.e., whence $x \equiv 0$. The corresponding cost is now $J = 0$.

F. Clarke, *Functional Analysis, Calculus of Variations and Optimal Control*,
Graduate Texts in Mathematics 264, DOI 10.1007/978-1-4471-4820-3_23,

The reader is to understand, therefore, that an arbitrarily small violation of the endpoint constraint (which one would probably tolerate in practice) leads to a significant, disproportionate decrease in the minimum of the problem. In other terms, if we define the value function $V(\beta)$ as the minimum in the problem when the endpoint constraint is given by $z(1) = \beta$ (other things being equal), the function V fails to be lower semicontinuous at 0: we have $V(0) = 1$, yet $V(\varepsilon) \leqslant 0$ for every $\varepsilon > 0$. □

Just as an unstable equilibrium in mechanics is considered somewhat meaningless, one may consider that the problem above is not well posed, since its solution lacks stability, in a certain sense. It is the goal of *relaxation* to reformulate the problem so as to avoid this phenomenon. Relaxation is closely related to the existence question, but it is a new consideration. Note that nonexistence of a solution was *not* the issue in the example above.

The pathology illustrated by Example 23.1 can be ascribed to a specific defect of the system: the set of state trajectories fails to be closed. This property is an essential one in existence theory, for evident reasons: in applying the direct method, one produces solutions by taking limits.

It turns out that the main property that a system (f,U) must possess to render the set of its trajectories closed is that the set $f(x,U)$ of available velocities be convex for each x. (This fact is foreshadowed by Exer. 8.45, which reveals the connection between weak closure and convexity.) Here is a basic sequential compactness theorem for system trajectories:

23.2 Theorem. *Let (f,U) be a control system on the interval $[a,b]$ for which:*

(a) *$f(t,x,u)$ is continuous in (x,u) and measurable in t;*

(b) *$U(\cdot)$ is measurable and compact valued;*

(c) *f has linear growth: there is a summable function M such that*

$$(t,x) \in [a,b] \times \mathbb{R}^n,\ u \in U(t) \implies |f(t,x,u)| \leqslant M(t)(1+|x|);$$

(d) *The set $f(t,x,U(t))$ is convex for each (t,x).*

Let (x_i,u_i) be a sequence of processes for the control system (f,U) such that the set $\{x_i(a) : i \geqslant 1\}$ is bounded. Then there exists a subsequence of x_i converging uniformly to a state trajectory x_* of the system.

The conclusion means that the limit x_* is an arc which admits a control function u_* satisfying

$$x_*'(t) = f\big(t, x_*(t), u_*(t)\big),\ t \in [a,b] \text{ a.e.}$$

Note, however, the total lack of any assertion that u_i converges in any particular sense to u_*.

Proof. Define a multifunction by $\Gamma(t,x) = f(t,x,U(t))$, which is a mapping from $[a,b]\times\mathbb{R}^n$ to the subsets of $\mathbb{R}^n$. Then the x_i satisfy the differential inclusion

$$x_i'(t) \in \Gamma\big(t, x_i(t)\big) \text{ a.e.}$$

We prepare an appeal to the weak closure theorem 6.39. The linear growth hypothesis of the theorem leads to a subsequence of the x_i (we do not relabel) converging uniformly to an arc x_*, and such that x_i' converges weakly in $L^1(a,b)$ to x_*' (we have seen this in Exer. 6.42).

The hypotheses imply that Γ is convex-valued, and that, for each t, the graph of $\Gamma(t,\cdot)$ is closed (exercise). Let S be a compact subset of $[a,b]\times\mathbb{R}^n$ containing the graphs of all the functions x_i. Then, for a certain constant K, for all $(t,x) \in S$, the set $\Gamma(t,x)$ is bounded by $KM(t)$, by the linear growth condition. This observation supplies hypothesis (c) of Theorem 6.39; there remains (b) to verify. This requires that the support function map

$$t \mapsto H_{\Gamma(t,y(t))}(p)$$

be measurable for every $p \in \mathbb{R}^n$ and measurable function $y(t)$. But since f is continuous in u, this function coincides with

$$t \mapsto \sup_{i \geqslant 1} \langle p, f(t,y(t),\gamma_i(t))\rangle,$$

where $\{\gamma_i\}$ is a countable family of measurable functions generating U as in Theorem 6.22. This confirms the required measurability: the map $t \mapsto f(t,y(t),\gamma_i(t))$ is measurable for each i (see Props. 6.35 and 6.34), and the upper envelope of countably many measurable functions is measurable.

By Theorem 6.39, the limit arc x_* satisfies the differential inclusion $x' \in \Gamma(t,x)$ a.e. The final step in the proof is to show that x_* is indeed the state component of a process (x_*,u_*). To this end, let

$$W(t) = \big\{u \in U(t) : |x_*'(t) - f(t,x_*(t),u)| = 0\big\}.$$

Then W is nonempty for almost every $t \in [a,b]$, and is a measurable multifunction by Prop. 6.25. Since the hypotheses imply that W is closed-valued, it follows from Cor. 6.23 that W admits a measurable selection $u(\cdot)$ on $[a,b]$, which confirms that x_* is a state trajectory. □

Filippov's lemma. The proof of Theorem 23.2 actually establishes that (under the hypotheses in force) the standard control system

$$x'(t) = f\big(t,x(t),u(t)\big),\ u(t) \in U(t),\ t \in [a,b] \text{ a.e.} \qquad (*)$$

is equivalent to the differential inclusion

$$x'(t) \in f\big(t,x(t),U(t)\big),\ t \in [a,b] \text{ a.e.} \qquad (**)$$

The term "equivalent" here means that the trajectories x of the two systems coincide. Now, it is clear that a state trajectory x of the control system $(*)$ is an arc that satisfies $(**)$; that is, a trajectory of the differential inclusion. It is the opposite implication that is nontrivial, and that can only be affirmed under certain hypotheses.

Given an arc x that satisfies $(**)$, the axiom of choice tells us that there is a function $u(t)$ satisfying $u(t) \in U(t)$ for every t, and such that $(*)$ holds; the issue is whether there is a *measurable* function u doing this. In many cases, the measurable selection theory of §6.2 is adequate to the task of verifying this, as it was in the proof of Theorem 23.2. The equivalence of $(*)$ and $(**)$ in that setting is known as *Filippov's lemma*, a term that is now used more generally for any result affirming this equivalence, even under weaker hypotheses.

When f is less regular than in Theorem 23.2, and when U is not necessarily closed-valued, more sophisticated measurable selection results are required to obtain corresponding extensions of Filippov's lemma. The concept of LB measurability plays a role in this endeavor.

Recall that a multifunction Γ from $\mathbb{R}^m$ to $\mathbb{R}^n$ is said to be LB measurable if its graph, the set

$$\operatorname{gr}\Gamma = \big\{(x,u) \in \mathbb{R}^m \times \mathbb{R}^n : u \in \Gamma(x)\big\},$$

belongs to the σ-algebra $L \times B$ (see Def. 6.33). When the values of Γ are closed sets, it can be shown that the LB measurability of Γ is equivalent to measurability as defined in §6.2. We admit this without proof, as well the following selection theorem that extends Cor. 6.23.[1] (Recall that the domain of Γ, denoted by $\operatorname{dom}\Gamma$, is the set of points x for which $\Gamma(x) \neq \emptyset$.)

23.3 Theorem. (Aumann's selection theorem) *If Γ is LB measurable, then there exists a measurable selection for Γ; that is, a Lebesgue measurable function γ mapping $\operatorname{dom}\Gamma$ to $\mathbb{R}^n$ such that $\gamma(x) \in \Gamma(x)$ for almost every $x \in \operatorname{dom}\Gamma$.*

This selection theorem leads to the following more refined version of Filippov's lemma.

23.4 Corollary. *Let the control system (f,U) be such that $f(t,x,u)$ is continuous in x for each (t,u) and LB measurable in (t,u) for each x, and such that $U(\cdot)$ is LB measurable. Then $(*)$ and $(**)$ have the same trajectories.*

Proof. Let x be an arc satisfying $(**)$. Then the issue is to produce a measurable selection of the multifunction

$$\Gamma(t) = \big\{u \in U(t) : x'(t) = f(t,x(t),u)\big\}.$$

[1] This advanced selection theorem is an outgrowth of the theory of Souslin sets; see [25].

The hypotheses imply that Γ is nonempty on $[a,b]$ a.e. Because the function $(t,u) \mapsto f(t,x(t),u)$ is LB measurable (by Prop. 6.36), we find that the set

$$\operatorname{gr}\Gamma = \{(t,u) \in \operatorname{gr} U : x'(t) - f(t,x(t),u) = 0\}$$

lies in $L \times B$, as the intersection of two sets in $L \times B$; thus, the required selection exists by Theorem 23.3. □

Relaxed trajectories. It is considered a positive feature if the system (f,U) happens to satisfy the hypotheses of Theorem 23.2, for then the state trajectories are sequentially compact, in the indicated sense. Besides being a principal factor in obtaining the existence of solutions in optimal control, this sequential compactness property precludes the ill-posedness pathology that we encountered in Example 23.1: in its presence, the function V discussed there is lower semicontinuous (as the reader may care to verify).

If (f,U) fails to have the indicated properties, it is likely to be the convexity of the velocity sets that goes wrong. Accordingly, it is natural to convexify those sets, an idea that leads to the following:

23.5 Definition. *A* **relaxed trajectory** *of the system* (f,U) *on an interval* $[a,b]$ *refers to an arc* y *which satisfies the differential inclusion*

$$y'(t) \in \overline{\operatorname{co}}\, f\big(t,y(t),U(t)\big),\ t \in [a,b] \text{ a.e.}$$

Note that the set of relaxed trajectories includes the set of usual ones, which are called *original trajectories* for the purposes of contrast.[2]

23.6 Exercise. For $n = 1$, let $f(x,U) = \{-1,+1\}\ \forall x$, and let y be a relaxed trajectory. Show that, for any $\varepsilon > 0$, there exists an original trajectory x of the system (f,U) such that

$$x(a) = y(a),\ |x(t) - y(t)| \leqslant \varepsilon\ \forall t \in [a,b]. \qquad \square$$

We introduced relaxed trajectories in the hope of achieving sequential compactness. Under the right assumptions, this has succeeded:

23.7 Exercise. Let the system (f,U) satisfy properties (a), (b), and (c) of Theorem 23.2. Prove that the relaxed trajectories are sequentially compact in the following sense: if y_i is a sequence of relaxed trajectories for which the set $\{y_i(a) : i \geqslant 1\}$ is bounded, then there exists a subsequence of y_i converging uniformly to a relaxed trajectory y of the system. □

[2] We are sidestepping the question of finding an explicit control system generating the relaxed trajectories. This is most often accomplished by introducing controls whose values are probability measures on the control set.

Under the hypotheses of the exercise above, the relaxed problem will *not* suffer from the pathology exhibited in Example 23.1. Furthermore, the extension to relaxed trajectories turns out to be a faithful one in general, in the sense that, when f is Lipschitz with respect to x, then the relaxed trajectories constitute precisely the *closure* of the set of original trajectories (see Exer. 26.18).

There are those who have concluded from these facts that only relaxed problems of optimal control should ever be considered. Of course, if the original system is *already* relaxed, in the sense that original and relaxed trajectories coincide, there is no cause for disagreement on the matter. A widely-studied class of control systems whose properties are relevant in this regard is the following.

23.8 Definition. *The control system* (f,U) *is said to be* **finitely generated** *if* f *has the form*

$$f(t,x,u) = g_0(t,x) + G(t,x)\,u = g_0(t,x) + \sum_{j=1}^{m} g_j(t,x)\,u^j,$$

where G *is a function whose values are* $n \times m$ *matrices*[3] *(whose columns are the vectors* $g_1, g_2, \dots, g_m$*), and* $u = (u^1, u^2, \dots, u^m) \in \mathbb{R}^m$*.*

Thus, a finitely generated system corresponds to (certain) linear combinations of a finite family of vector fields $\{g_i : 1 \leqslant i \leqslant m\}$, added to the *drift* term g_0. When f has this form, we also say that f is *affine in the control variable*.

23.9 Exercise. We consider a system (f,U) which is finitely generated, and where each of the vector fields g_j $(j = 0,1,\dots,m)$ is measurable in t, continuous in x, and has linear growth, as follows: there is a summable function M such that

$$(t,x) \in [a,b] \times \mathbb{R}^n \implies |g_j(t,x)| \leqslant M(t)(1+|x|).$$

Suppose in addition that the control set U is a compact convex set not depending on t. Prove that under these assumptions, the system is relaxed (original and relaxed trajectories coincide), and that it satisfies the hypotheses of Theorem 23.2. □

23.2 Three existence theorems

Let us contemplate what factors should play a role in deriving general existence theorems in optimal control. To begin with, and roughly speaking, such results require that velocity sets be convex, for reasons which should now be apparent following our discussion of relaxation. As for the running cost Λ, its convexity relative to the

[3] We continue to adhere to the convention that points in Euclidean space, in their dealings with matrices, take the form of columns.

control variable is the functional counterpart of that property. It is natural as well to impose linear growth on the dynamics (to prevent finite-time blowup, see p. 260). In addition, a growth condition on controls will be required. This is most easily supplied by taking the control set U to be compact; otherwise, when the controls are unbounded, coercivity (in u) of the running cost can be postulated (this hearkens back to Tonelli's theorem). The final ingredient is the lower semicontinuity of the boundary cost ℓ and the closedness of the target set.

There is no mystery as to why these factors play a role in existence theory: they are the ones required to apply the direct method, studied in detail in Part III. The reader will recall our philosophy: mastering the underlying method is preferable to depending entirely on pre-formulated existence theorems. Nonetheless, these can save time, so we give three representative ones that apply to different scenarios.

Existence for the Mayer problem. The first of our three existence theorems concerns the *Mayer form* of the optimal control problem: it is characterized by a cost that depends only upon the endpoint values of the state. Specifically, we consider the following problem on a fixed underlying interval $[a,b]$:

$$\left\{\begin{array}{ll} \text{Minimize} & J(x,u) = \ell\big(x(a),x(b)\big) \\ \text{subject to} & x'(t) = f\big(t,x(t),u(t)\big) \text{ a.e.} \\ & u(t) \in U(t) \text{ a.e.} \\ & \big(t,x(t)\big) \in Q \;\forall\, t \in [a,b],\; \big(x(a),x(b)\big) \in E. \end{array}\right. \qquad \textbf{(OC1)}$$

Note that we have included a *unilateral state constraint* $(t,x(t)) \in Q$ in the formulation, where Q is a given subset of $[a,b]\times\mathbb{R}^n$. Such constraints often serve to localize the problem and the hypotheses.

23.10 Theorem. *Let the data of* (OC1) *satisfy the following hypotheses*:

(a) $f(t,x,u)$ *is continuous in* (x,u) *and measurable in* t;

(b) $U(\cdot)$ *is measurable and compact-valued*;

(c) f *has linear growth on* Q: *there is a summable function* M *such that*

$$(t,x) \in Q,\; u \in U(t) \implies |f(t,x,u)| \leqslant M(t)(1+|x|);$$

(d) *For each* $(t,x) \in Q$, *the set* $f(t,x,U(t))$ *is convex*;

(e) *The sets* Q *and* E *are closed, and* $\ell : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ *is lower semicontinuous*;

(f) *The following set is bounded*:

$$\{\alpha \in \mathbb{R}^n : (\alpha,\beta) \in E \text{ for some } \beta \in \mathbb{R}^n\}.$$

Then, if there is at least one admissible process for the problem, it admits a solution.

Proof. Fellow aficionados of the direct method (such as the reader) will recognize the familiar lines of the proof. We observe first that there exists a minimizing sequence by assumption.

According to Theorem 23.2 (adapted in the evident way to the presence of the set Q), we have sequential compactness: any minimizing sequence (x_i, u_i) necessarily admits a subsequence (we do not relabel) such that the arcs x_i converge uniformly to a state trajectory x_* of the control system. Since E and Q are closed, it follows that x_* is admissible for (OC1). The lower semicontinuity of ℓ yields

$$\ell\big(x_*(a), x_*(b)\big) \leqslant \lim_{i\to\infty} \ell\big(x_i(a), x_i(b)\big) = \inf\,(\text{OC1}).$$

Letting u_* be a control function for x_* (Filippov's lemma!), it follows that (x_*, u_*) is an optimal process. □

Existence with a running cost. When a running cost $\Lambda(t,x,u)$ which depends on u is present (in contrast to the Mayer problem above), the existence issue is more complex, for reasons we proceed to explain.

In the presence of the running cost, the use of the direct method must involve hypotheses which have the effect of imposing weak sequential compactness of the controls along a minimizing sequence (not just compactness of the state trajectories), as well as lower semicontinuity of the integral functional with respect to this convergence.

We have seen how to do this in prior work on integral functionals in the calculus of variations. The coercivity of Λ would lead to the weak compactness, but now we can also exploit the (possible) compactness of the control set. This explains the alternative hypotheses in the next theorem.

A final point concerns the structure of f. In contrast to the Mayer case, the argument now involves a weak limit of a sequence of controls, for the reasons given above. In Theorem 23.10, the limiting control is generated *after* the convergence of the state trajectories, by Filippov's lemma. This approach is unavailable now, and in order for the weak limit to preserve the dynamics, these must be affine with respect to u. This explains why, below, we have made the underlying system a finitely generated one (see Def. 23.8).

We consider the following optimal control problem defined on a fixed underlying interval $[a,b]$.

$$\begin{cases} \text{Minimize} & J(x,u) = \ell\big(x(a), x(b)\big) + \displaystyle\int_a^b \Lambda\big(t, x(t), u(t)\big)\,dt \\ \text{subject to} & x'(t) = g_0\big(t, x(t)\big) + \textstyle\sum_{j=1}^m g_j\big(t, x(t)\big)\,u^j(t) \text{ a.e.} \\ & u(t) \in U(t) \text{ a.e.} \\ & \big(t, x(t)\big) \in Q \;\; \forall\, t \in [a,b], \;\; \big(x(a), x(b)\big) \in E. \end{cases} \qquad \textbf{(OC2)}$$

The existence theorem has a rather long statement that goes as follows:

23.11 Theorem. *Let the data of* (OC2) *satisfy the following hypotheses*:

(a) *Each* g_j $(j = 0,1,\ldots,m)$ *is measurable in* t, *continuous in* x, *and has linear growth*: *there exists a constant* M *such that*

$$(t,x) \in Q \implies |g_j(t,x)| \leqslant M(1+|x|);$$

(b) *For almost every* t, *the set* $U(t)$ *is closed and convex*;

(c) *The sets* E *and* Q *are closed, and* $\ell : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ *is lower semicontinuous*;

(d) *The running cost* $\Lambda(t,x,u)$ *is* LB *measurable in* t *and* (x,u), *and lower semicontinuous in* (x,u); $\Lambda(t,x,\cdot)$ *is convex for each* $(t,x) \in Q$; *there is a constant* λ_0 *such that*

$$(t,x) \in Q,\ u \in U(t) \implies \Lambda(t,x,u) \geqslant \lambda_0 .$$

(e) *The projection* $\{\alpha \in \mathbb{R}^n : (\alpha,\beta) \in E \text{ for some } \beta \in \mathbb{R}^n\}$ *of* E *is bounded*;

(f) *One of the following holds for some* $r > 1$:

(i) *There exists* $k \in L^r(a,b)$ *such that, for almost every* t,

$$u \in U(t) \implies |u| \leqslant k(t), \text{ or}$$

(ii) *There exist* $\alpha > 0$ *and* β *such that*

$$(t,x) \in Q,\ u \in U(t) \implies \Lambda(t,x,u) \geqslant \alpha|u|^r + \beta .$$

Then, if there is at least one admissible process (x,u) *for which* $J(x,u)$ *is finite, the problem admits a solution.*

Proof. Let (x_i,u_i) be a minimizing sequence. The following analysis will seem rather familiar, from the use of the direct method in the calculus of variations. It is easy to see that either alternative in hypothesis (f) implies that the sequence u_i is bounded in $L^r(a,b)$. Since this space is reflexive (Theorem 6.4), we can invoke weak sequential compactness (Theorem 5.50) and assume (for a subsequence, without relabeling) that u_i converges weakly in $L^r(a,b)$ to a limit u_*. It follows from Hölder's inequality that the sequence u_i is also bounded in $L^1(a,b)$.

Using (a), we obtain an estimate of the following type for the sequence x_i:

$$|x_i'(t)| \leqslant M\big(1+|x_i(t)|\big)\big(1+K|u_i(t)|\big) \text{ a.e.}$$

With the help of Gronwall's lemma (Theorem 6.41), because u_i is bounded in $L^1(a,b)$, and in light of (e), we deduce from this estimate that the sequence x_i is uniformly bounded on $[a,b]$. Returning to the estimate, it now follows that x_i' is bounded in $L^r(a,b)$.

This allows us to invoke weak sequential compactness in $L^r(a,b)$ once more, now for the sequence x_i'. Thus we may suppose that x_i' converges weakly to a limit v_* in $L^r(a,b)$. We also derive the equicontinuity of the sequence x_i, by Hölder's inequality. Calling upon Ascoli's theorem (and continuing to take subsequences), there is a continuous function x_* which is the uniform limit of the x_i. It follows from the identity

$$x_i(t) = x_i(a) + \int_a^t x_i'(s)\,ds$$

(by taking limits) that x_* is an arc and that $x_*'(t) = v_*(t)$ a.e.

We may summarize our conclusions to this point as follows:

u_i converges weakly in $L^r(a,b)$ to u_*, x_i converges uniformly to an arc x_*, and x_i' converges weakly to x_*' in $L^r(a,b)$.

For the next step, let us note that the problem is unchanged if Λ is replaced by $\max(\Lambda, \lambda_0)$, in view of hypothesis (d); this preserves convexity and lower semicontinuity. Then we may invoke the integral semicontinuity theorem 6.38 (and the lower semicontinuity of ℓ) to deduce that

$$J(x_*, u_*) \leqslant \liminf_{i\to\infty} J(x_i, u_i) = \inf(\text{OC}\,2).$$

It follows that (x_*, u_*) solves (OC 2), provided of course that (x_*, u_*) is an admissible process for the problem. We proceed now to prove this.

The state and boundary constraints are preserved in the limit, since E and Q are closed (hypothesis (c)). The set

$$W = \{ w \in L^r(a,b) : w(t) \in U(t) \text{ a.e.} \}$$

is strongly closed in $L^r(a,b)$, since strongly convergent sequences admit a subsequence converging almost everywhere, and since $U(\cdot)$ is closed-valued by hypothesis (b). W is also convex (again by (b)), and therefore W is weakly closed (Theorem 3.6). It follows that u_* (as the weak limit of the sequence u_i) belongs to W, and is therefore a control function.

There remains only to verify that x_* is the state trajectory corresponding to u_*. To do this, it suffices to show that, for any measurable subset A of $[a,b]$, we have

$$\int_A \Big\{ x_*'(t) - g_0\big(t, x_*(t)\big) - \textstyle\sum_{j=1}^m g_j\big(t, x_*(t)\big)\, u_*^j \Big\}\, dt = 0.$$

This equality holds when x_* and u_* are replaced by x_i and u_i respectively. To obtain the desired conclusion, it suffices to justify passing to the limit as $i \to \infty$. By weak convergence, and by the dominated convergence theorem, we have

$$\int_A x_i'(t)\,dt \to \int_A x_*'(t)\,dt \quad\text{and}\quad \int_A g_0\big(t, x_i(t)\big)\,dt \to \int_A g_0\big(t, x_*(t)\big)\,dt.$$

We also know that for each $j \in \{1,2,\dots,m\}$, the following holds:

$$\int_A g_j\big(t,x_*(t)\big)\,u_i^j(t)\,dt \;\to\; \int_A g_j\big(t,x_*(t)\big)\,u_*^j(t)\,dt \;\text{ as } i \to \infty,$$

since $g_j(t,x_*(t)) \in L^\infty(a,b)$ by hypothesis (a). To complete the proof, it suffices to show that

$$\int_A \Big\{ g_j\big(t,x_i(t)\big) - g_j\big(t,x_*(t)\big) \Big\}\, u_i^j(t)\,dt \;\to\; 0\,.$$

By Hölder's inequality, the integral on the left is bounded in absolute value by

$$\Big\{ \int_a^b \big|g_j\big(t,x_i(t)\big) - g_j\big(t,x_*(t)\big)\big|^{r_*}\,dt \Big\}^{1/r_*} \big\|\,u_i^j\,\big\|_{L^r(a,b)}\,.$$

The first factor in this product tends to 0, by dominated convergence, and the second is bounded independently of i since the sequence u_i is bounded in $L^r(a,b)$. The result follows. □

Remark. In practice, the state constraint involving the set Q is often added to the problem in order to be able to invoke existence theory, but subsequently, one would prefer to discard it if possible, in order to facilitate the writing of necessary conditions. As an example of this procedure, consider the problem of §22.2. The state constraint $x(t) \geqslant 0$ is automatically satisfied because of the nature of the problem (as is easily seen), so that one may add to it the constraint

$$(t,x(t)) \in Q := [\,0,T\,] \times \mathbb{R}_+$$

without actually changing anything. However, this modification allows us to verify the existence of λ_0 as required by hypothesis (d) of Theorem 23.11. It is clear that the other hypotheses are present, so we are able to assert that a solution exists. Subsequently, in order to be able to apply the maximum principle, we may once again safely ignore the constraint $x(t) \geqslant 0$, since it is automatically satisfied (as we have said).

23.12 Exercise. For a positive constant ρ, we consider the problem

$$\min \int_0^1 x'(t)^3\,dt\,,\; x \in \text{Lip}[0,1]\,,\; |x'(t)-1| \leqslant \rho\,,\; x(0)=0,\; x(1)=1\,.$$

This is the problem studied in Example 15.15, with an extra constraint on the velocity. It can be viewed as an optimal control problem in the evident way, by identifying the control u with x', and by taking the control set U to be the interval $[\,1-\rho\,,1+\rho\,]$.

(a) Prove that if $\rho \leqslant 1$, then a solution x_* exists.

(b) Use the maximum principle to show that $x_*(t) = t$.

(c) Prove that for $\rho > 3$, the arc x_* is *not* a solution of the problem. □

Existence with variable time. Our third and last existence theorem for optimal control treats a variable-time problem with finitely generated dynamics. We consider the problem

$$\begin{cases} \text{Minimize} & J(\tau,x,u) = \int_0^{\tau} \Lambda\big(x(t),u(t)\big)\,dt \\ \text{subject to} & \tau \geqslant 0 \\ & x'(t) = f\big(x(t),u(t)\big) \\ & \quad = g_0\big(x(t)\big) + \sum_{j=1}^m g_j\big(x(t)\big)u^j(t) \text{ a.e.} \\ & u(t) \in U \text{ a.e.} \\ & x(t) \in S \;\; \forall t \in [0,\tau]. \\ & x(0) = x_0,\; x(\tau) \in E. \end{cases} \tag{OC3}$$

The corresponding existence result is the following:

23.13 Theorem. *Let the data of* (OC3) *satisfy the following hypotheses*:

(a) *Each* g_j $(j = 0, 1, \ldots, m)$ *is continuous and has linear growth*: *there is a constant* M *such that*

$$x \in S \implies |g_j(x)| \leqslant M(1+|x|);$$

(b) U *is compact and convex*;

(c) E *and* S *are closed subsets of* $\mathbb{R}^n$, *with* $S \supset E$;

(d) $\Lambda(x,u)$ *is lower semicontinuous in* (x,u), *convex in* u, *and bounded below by a constant* $\lambda_0 > 0$ *as follows*: $x \in S,\ u \in U \implies \Lambda(x,u) \geqslant \lambda_0$;

(e) *For every* $x \in E$, *we have* $T_E(x) \cap f(x,U) \neq \emptyset$.

Then, if there is at least one admissible process which joins x_0 *to* E *in finite time and for which the cost is finite, the problem admits a solution.*

We recognize the last hypothesis, which involves the tangent cone to E, as saying that the system $(E, f(x,U))$ is weakly invariant, by Theorem 12.7. This has a natural interpretation: once we attain the target, we want to be able to stay there. Note that in the minimal-time problem we have $\Lambda \equiv 1$, which satisfies the hypotheses imposed on the running cost.

Proof. Let $(\bar{x}, \bar{u})$ be an admissible process which joins x_0 to E in finite time $\bar{\tau}$, and for which $J(\bar{\tau}, \bar{x}, \bar{u})$ is finite. The minimization in (OC 3) may be restricted to admissible processes (x,u) defined on an interval $[0,\tau]$ for which

$$J(\bar{\tau},\bar{x},\bar{u}) \geqslant J(\tau,x,u) = \int_0^{\tau} \Lambda\big(x(t),u(t)\big)\,dt \geqslant \lambda_0\,\tau.$$

In particular, we may limit to $\tau \leqslant J(\bar{\tau},\bar{x},\bar{u})/\lambda_0 =: T$. Note that $T \geqslant \bar{\tau}$.

We define a new running cost by

$$\Lambda_+(x,u) = \begin{cases} \Lambda(x,u) & \text{if } x \notin E \\ 0 & \text{otherwise.} \end{cases}$$

We verify without difficulty that Λ_+ is lower semicontinuous (and therefore Borel measurable), and convex in u.

We now consider the fixed-time problem (P) of minimizing

$$J_+(x,u) = \int_0^T \Lambda_+\big(x(t),u(t)\big)\,dt$$

subject to the given dynamics, and with the state and boundary constraints

$$x(t) \in S \;\; \forall t \in [0,T], \;\; x(0) = x_0, \;\; x(T) \in E.$$

We may extend $\bar{x}$ to $[\bar{\tau},T]$ in such a way that it remains in E (and hence in S), by the weak invariance cited above. (Because of linear growth, Theorem 12.3 applies: the extension can be defined on $[\bar{\tau},\infty)$.) This produces a process admissible for (P) and having finite cost. A direct application of Theorem 23.11 shows that (P) admits a solution (x_*,u_*).

Let $\tau_* \leqslant T$ be the first time such that $x_*(\tau_*) \in E$. Then $x_*(t) \in E \;\; \forall t \in [\tau_*,T]$, or else we could redefine x_* to remain in E on that interval, and, in so doing, obtain a strictly lower value of J_+, a contradiction.

We claim that the process (x_*,u_*), truncated to $[0,\tau_*]$, solves (OC 3). If this fails to be the case, there exists a process (x,u) on some interval $[0,\tau]$ such that

$$J(\tau,x,u) < J(\tau_*,x_*,u_*) = J_+(x_*,u_*) \leqslant J_+(\bar{x},\bar{u}),$$

by the optimality of (x_*,u_*). As before, this implies $\tau < T$. Now extend x to $[\tau,T]$ so that it remains in E; then we obtain a process (x,u), admissible for (P), which satisfies $J_+(x,u) = J(\tau,x,u) < J_+(x_*,u_*)$: a contradiction. □

23.14 Exercise. We now ask the reader to return to certain examples discussed earlier, in order to verify that some claims made at the time, in regard to the existence of solutions, were justified.

(a) Consider the problem of Example 22.9. Show that the constraint $|x(t)| \leqslant 4$ may be adjoined to it, without really changing anything. Then observe that, with this modification, Theorem 23.11 becomes applicable. Deduce that a solution exists.

(b) Show that the soft landing problem (Example 22.14) admits a solution.

(c) Show that the problem (LQR) of Example 22.19 admits a solution.

(d) Show that the problem of Example 22.28 admits a solution. □

23.3 Regularity of optimal controls

It is possible for optimal controls to be very discontinuous, even when the data of the problem are very smooth.

23.15 Example. Let $u_* : [0,1] \to \mathbb{R}^n$ be *any* measurable function with values in the unit ball in $\mathbb{R}^n$, and define $x_* : [0,1] \to \mathbb{R}^n$ by

$$x_*(t) = \int_0^t u_*(s)\,ds.$$

Then x_* is Lipschitz continuous, so that its graph G is a closed subset of $[0,1]\times\mathbb{R}^n$. By a well-known result in topology (not hard to prove) there exists a nonnegative C^∞ function $\Lambda : \mathbb{R}^{n+1} \to \mathbb{R}$ such that the set $\{(t,x) : \Lambda(t,x) = 0\}$ is precisely the set G. Consider now the following optimal control problem, which has C^∞ data, a free endpoint, and a running cost independent of the control:

$$\begin{cases} \text{Minimize} & J(x,u) = \displaystyle\int_0^1 \Lambda\big(t, x(t)\big)\,dt \\ \text{subject to} & x'(t) = u(t),\ t \in [0,1] \text{ a.e.} \\ & u(t) \in U = B(0,1),\ t \in [0,1] \text{ a.e.} \\ & x(0) = 0. \end{cases}$$

It is clear that the unique optimal process (the only one giving zero cost) is (x_*, u_*). The optimal control u_* is as discontinuous as we like (but measurable); in particular, it can fail to be piecewise continuous. □

We have seen how convexity of the running cost $\Lambda(t,x,u)$ with respect to u plays a crucial role in existence theory. For purposes of deducing the regularity of optimal controls, a certain strengthening of that property is useful. We say that the running cost Λ is **strongly convex** in u if, for every bounded subset C of $[a,b]\times\mathbb{R}^n\times\mathbb{R}^m$, there exists $c > 0$ such that

$$(t,x,u),\,(t,x,v) \in C \implies \langle \Lambda_u(t,x,v) - \Lambda_u(t,x,u), v-u \rangle \geqslant c\,|v-u|^2.$$

The reader will notice that we suppose, in writing this, that Λ is differentiable with respect to u. We may think of strong convexity as a calibrated form of strict convexity, as the following exercise shows.

23.16 Exercise.

(a) Prove that if Λ is strongly convex in u, then the mapping $u \mapsto \Lambda(t,x,u)$ is strictly convex on $\mathbb{R}^m$ for each (t,x).

(b) Suppose that Λ is twice continuously differentiable with respect to u, and that the Hessian matrix $D_u^2\Lambda$ is uniformly positive definite on bounded sets, in the

following sense: for every bounded subset C of $[a,b]\times\mathbb{R}^n\times\mathbb{R}^m$, there exists $c>0$ such that

$$\langle D_u^2\Lambda(t,x,u)w,w\rangle \geqslant c|w|^2 \ \ \forall w\in\mathbb{R}^m,\ \ \forall (t,x,u)\in C.$$

Prove that Λ is strongly convex.

(c) Let Λ be of the form $\Lambda(t,x,u) = \Lambda_0(t,x)+\langle Mu,u\rangle$, where the $m\times m$ matrix M is positive definite. Prove that Λ is strongly convex. □

We consider now the following optimal control problem, defined on a fixed underlying interval $[a,b]$, in which the system is finitely generated:

$$\left\{\begin{array}{ll} \text{Minimize} & J(x,u) = \ell\big(x(a),x(b)\big)+\int_a^b \Lambda\big(t,x(t),u(t)\big)\,dt \\ \text{subject to} & x'(t) = g_0\big(t,x(t)\big) + \sum_{j=1}^m g_j\big(t,x(t)\big)u^j(t) \text{ a.e.} \\ & u(t)\in U \text{ a.e.} \\ & \big(x(a),x(b)\big)\in E. \end{array}\right. \qquad \textbf{(OC4)}$$

The affine structure of the dynamics lends itself to proving the regularity of optimal controls when the running cost is strictly convex in the control variable, by exploiting the conclusions of the maximum principle. (Sometimes, in non affine settings, this can be done on an *ad hoc* basis.) The following result is foreshadowed (of course) in the calculus of variations (see Theorem 15.5).

23.17 Theorem. *Let* (x_*,u_*) *be an admissible process for* (OC4) *which satisfies the necessary conditions of the extended maximum principle* (Theorem 22.26) *in the normal case* ($\eta=1$). *Suppose that the following hypotheses hold*:

(a) *The functions* ℓ, Λ, *and* g_j $(j=0,\dots,m)$ *are locally Lipschitz*;

(b) U *is a compact convex subset of* $\mathbb{R}^m$;

(c) *For each* (t,x), *the map* $u\mapsto\Lambda(t,x,u)$ *is strictly convex.*

Then u_* *is continuous. If, in addition, the following holds*:

(d) Λ *is differentiable with respect to* u, Λ_u *is locally Lipschitz, and* Λ *is strongly convex in* u,

then u_* *is Lipschitz continuous.*

As usual in dealing with measurable functions, the statement "u_* is continuous" means that there is a continuous function that agrees with u_* a.e.; we also say that u_* admits a continuous representative.

Proof. The hypotheses imply that u_* is essentially bounded and (since x_*' is bounded) that x_* is Lipschitz. Let p be the costate of Theorem 22.26. It follows

that for a certain constant K, for almost every t, the function

$$x \mapsto H\big(t, x, p(t), u_*(t)\big) = \langle p(t), f(t, x, u_*(t)) \rangle - \Lambda(t, x, u_*(t))$$

is Lipschitz of rank K near $x_*(t)$. (Here, f denotes the dynamics function of the finitely generated system.) Then the generalized gradient

$$\partial_C H\big(t, \bullet, p(t), u_*(t)\big)(x_*(t))$$

is a subset of $B(0,K)$ (Prop. 10.5), so it follows from the adjoint inclusion (A) that the costate p has bounded derivative, and hence is Lipschitz.

As usual, we denote by G the $n \times m$ matrix whose columns are the elements g_j $(j = 1, 2 \ldots, m)$. The stationarity condition (Fermat's rule) corresponding to the maximum condition (M) asserts that (see Prop. 4.12) the costate p satisfies, for almost every t:

$$G^*\big(t, x_*(t)\big)p(t) \in \partial_u \Lambda\big(t, x_*(t), u_*(t)\big) + N_U\big(u_*(t)\big).$$

From this point, it is an easy matter to adapt the proof of Theorem 15.5 to show that there is a continuous function which agrees with u_* almost everywhere. The argument is essentially the same as before, and is based on the fact that if, for a given τ, two different points u_1, u_2 satisfy

$$G^*\big(\tau, x_*(\tau)\big)p(\tau) \in \partial_u \Lambda\big(\tau, x_*(\tau), u_i\big) + N_U\big(u_i\big),$$

then $u_1 = u_2$. This in turn follows from the fact that each u_i minimizes over the convex set U the strictly convex function

$$u \mapsto \Lambda\big(\tau, x_*(\tau), u\big) - \langle p(\tau), G(\tau, x_*(\tau))u \rangle.$$

Having dealt with the strictly convex case, we now turn to the proof of Lipschitz continuity, under hypothesis (d). We begin by defining two useful parameters. The first, denoted by L, is a common Lipschitz constant on $[a,b]$ for all the functions

$$r \mapsto G^*\big(r, x_*(r)\big)p(r), \quad r \mapsto \Lambda_u\big(r, x_*(r), u\big) \ \ (u \in U).$$

The existence of L follows from the fact that x_* and p are Lipschitz, while G and Λ_u are Lipschitz on bounded sets (by assumption).

The second parameter $c > 0$ is taken to be a strong convexity constant for Λ relative to a bounded set C large enough to contain $\text{gr}\,(x_*) \times U$.

Now let us fix any two values of s and $t \in [a,b]$ for which the maximum condition (M) holds. Then, for certain vectors n_s, n_t in $\mathbb{R}^m$, we have (by stationarity)

$$G^*\big(s, x_*(s)\big)p(s) - \Lambda_u\big(s, x_*(s), u_*(s)\big) = n_s \in N_U\big(u_*(s)\big), \tag{1}$$

$$G^*\big(t, x_*(t)\big)p(t) - \Lambda_u\big(t, x_*(t), u_*(t)\big) = n_t \in N_U\big(u_*(t)\big). \tag{2}$$

From the definition of a normal to a convex set, it follows that

$$\langle n_t - n_s, u_*(s) - u_*(t) \rangle \leqslant 0. \tag{3}$$

Invoking strong convexity, we calculate as follows:

$$\begin{aligned}
c\,|u_*(s) - u_*(t)|^2 &\leqslant \langle \Lambda_u\big(s, x_*(s), u_*(s)\big) - \Lambda_u\big(s, x_*(s), u_*(t)\big), u_*(s) - u_*(t) \rangle \\
&= \langle \Lambda_u\big(s, x_*(s), u_*(s)\big) - \Lambda_u\big(t, x_*(t), u_*(t)\big), u_*(s) - u_*(t) \rangle \\
&\quad + \langle \Lambda_u\big(t, x_*(t), u_*(t)\big) - \Lambda_u\big(s, x_*(s), u_*(t)\big), u_*(s) - u_*(t) \rangle \\
&= \langle G^*\big(s, x_*(s)\big)p(s) - n_s - G^*\big(t, x_*(t)\big)p(t) + n_t, u_*(s) - u_*(t) \rangle \\
&\quad + \langle \Lambda_u\big(t, x_*(t), u_*(t)\big) - \Lambda_u\big(s, x_*(s), u_*(t)\big), u_*(s) - u_*(t) \rangle \ \text{(by (1) (2))} \\
&\leqslant 2L\,|s - t|\,|u_*(s) - u_*(t)|,
\end{aligned}$$

by (3), and by the Lipschitz condition. It follows that

$$|u_*(s) - u_*(t)| \leqslant (2L/c)\,|s - t|.$$

Since s and t are any two points in a subset of $[a, b]$ of full measure, it follows that u_* has a representative that is Lipschitz of rank $2L/c$. □

23.18 Exercise. Show that Theorem 23.17 applies to exactly one of the following: Example 22.9, the problem of §22.2, Example 22.28. □

Chapter 24

Inductive methods

In seeking to solve an optimal control problem, it may happen that we suspect that we have identified the solution, but the deductive reasoning that would allow us to assert its optimality is unavailable. This might be the case because no existence theorem applies, or because the applicability of the necessary conditions is uncertain. In such a situation, we may seek to use an inductive method to confirm the optimality of the suspect. We describe three such methods in this chapter. The first of these is based on a strengthening of the conditions that appear in the maximum principle.

24.1 Sufficiency by the maximum principle

Let us consider the general optimal control problem

$$\begin{cases} \text{Minimize} & J(x,u) = \ell\big(x(a),x(b)\big) + \displaystyle\int_a^b \Lambda\big(t,x(t),u(t)\big)\,dt \\ \text{subject to} & x'(t) = f\big(t,x(t),u(t)\big),\ t \in [a,b] \text{ a.e.} \\ & u(t) \in U(t),\ t \in [a,b] \text{ a.e.} \\ & \big(x(a),x(b)\big) \in E. \end{cases} \tag{EC}$$

Suppose that, in seeking to solve (EC), we have identified an admissible process (x_*, u_*) that (for a certain costate p) satisfies all the necessary conditions provided by the extended maximum principle (see Theorem 22.26). It does *not* follow that (x_*, u_*) is a local minimizer in any sense, *cela va sans dire*: the necessary conditions are not sufficient. However, we now proceed to identify a special context in which a certain set of conditions of maximum principle type *do* turn out to be sufficient for optimality.

F. Clarke, *Functional Analysis, Calculus of Variations and Optimal Control*, Graduate Texts in Mathematics 264, DOI 10.1007/978-1-4471-4820-3_24,

We assume that the data satisfy the basic hypotheses of § 22.6 (see p. 464), as well as the local Lipschitz hypothesis 22.25. Recall that in the normal case ($\eta = 1$) of the maximum principle, the maximized Hamiltonian is given by

$$M(t,x,p) = \sup_{u \in U(t)} H(t,x,p,u) = \sup_{u \in U(t)} \langle p, f(t,x,u) \rangle - \Lambda(t,x,u).$$

24.1 Theorem. *Let* (x_*, u_*) *be an admissible process for problem* (EC), *where* E *and* ℓ *are convex. Suppose that there exists a costate arc* p *that satisfies* (*with* $\eta = 1$) *the transversality condition* (T) *and the maximum condition* (M) (*almost everywhere*) *of* Theorem 22.26, *as well as, for some* $\delta > 0$, *the following*:

$$M\big(t, x_*(t)+y, p(t)\big) - M\big(t, x_*(t), p(t)\big) \leqslant \langle -p'(t), y \rangle \quad \forall\, y \in B(0,\delta),\ t \in [a,b] \text{ a.e.} \qquad (\mathrm{A}^*)$$

Then (x_*, u_*) *is a minimizer for* (EC) *relative to* $\|x - x_*\| \leqslant \delta$.

Proof. Let (x,u) be any admissible process for (EC), with $\|x - x_*\| \leqslant \delta$. Then, almost everywhere, the expression

$$\langle p(t), f(t,x(t),u(t)) \rangle - \Lambda(t,x(t),u(t))$$

is bounded above by $M(t,x(t),u(t))$ (by definition); in turn, (A^*) and (M) together imply that this term is bounded above by

$$\langle p(t), f(t,x_*(t),u_*(t)) \rangle - \Lambda(t,x_*(t),u_*(t)) - \langle p'(t), x(t) - x_*(t) \rangle.$$

Let us substitute $f(t,x(t),u(t)) = x'(t)$ (and similarly for (x_*, u_*)) and integrate over $[a,b]$. After rearranging, we discover

$$\int_a^b \Lambda\big(t,x(t),u(t)\big)\,dt \geqslant \int_a^b \Lambda\big(t,x_*(t),u_*(t)\big)\,dt + \langle p(t), x(t) - x_*(t) \rangle \Big|_a^b . \qquad (1)$$

Because E and ℓ are convex, the transversality condition (T) is equivalent to

$$\big(p(a), -p(b)\big) \in \partial \ell\big(x_*(a), x_*(b)\big) + N_E\big(x_*(a), x_*(b)\big),$$

as follows from Prop. 11.12 and Theorem 11.36. In turn, this implies that the element $(p(a), -p(b))$ lies in the subdifferential of the convex function $g = \ell + I_E$, by Theorem 4.10 and Exer. 4.5. Then the subgradient inequality for g yields

$$\ell\big(x(a), x(b)\big) - \ell\big(x_*(a), x_*(b)\big) \geqslant \langle (p(a), -p(b)), \big(x(a) - x_*(a), x(b) - x_*(b)\big) \rangle. \qquad (2)$$

Combining the inequalities (1) and (2) reveals $J(x,u) \geqslant J(x_*, u_*)$, which proves the theorem. □

24.2 Corollary. *Suppose that for almost every t, the function*

$$x \mapsto H\big(t, x, p(t), u_*(t)\big)$$

is concave on $B(x_*(t), \delta)$. Then the conclusion of Theorem 24.1 *holds if the hypothesis* (A^*) *is replaced by the adjoint inclusion* (A) *of* Theorem 22.26.

Proof. It suffices to verify that condition (A^*) holds. When H has the stated concavity property, the function

$$\varphi(x) = (-H)\big(t, x, p(t), u_*(t)\big) + I_{B(x_*(t),\delta)}(x)$$

(for fixed t) is convex. In this case we may write (for almost every t), with the help of Theorem 10.8 and Prop. 10.11:

$$\partial \varphi\big(x_*(t)\big) = \partial_C\, \varphi\big(x_*(t)\big) = -\partial_C(-\varphi)\big(x_*(t)\big) \ni p'(t),$$

by (A). It now follows that (A^*) holds: it is simply the subgradient inequality at $x_*(t)$ corresponding to $p'(t)$ and the convex function φ (bearing in mind (M)). □

Remark. Note that no convexity of Λ is postulated in Theorem 24.1. It is clear, however, from the definition of H, that the concavity property cited in the corollary will hold if the dynamics of the problem are affine in the state variable and Λ is convex in x (exercise). In such cases, then, the maximum principle is rather close to being a sufficient, as well as a necessary, condition.

24.3 Example. We return to Example 18.16 in order to illustrate the use of Theorem 24.1. Recall that the problem consists of minimizing

$$\int_0^T \big\{ |x(t)| + g\big(|x'(t)|\big) \big\}\, dt$$

subject to the endpoint constraint $x(T) = \beta$, with $x(0)$ free, where g is given by

$$g(r) = \begin{cases} 1 + r^2/2 & \text{if } r \neq 0 \\ 0 & \text{if } r = 0. \end{cases}$$

In the case $0 < T \leqslant \sqrt{2}$, which is the only one we revisit, we had conjectured that the solution was $x_*(t) \equiv \beta$. We prove this now.

We recognize that the problem is a special case of (EC), with data

$$[a,b] = [0,T],\ E = \mathbb{R} \times \{\beta\},\ \ell = 0,\ f(t,x,u) = u,\ U(t) = \mathbb{R},\ \Lambda = |x| + g(|u|).$$

Thus, E and ℓ are convex. We have $H = p\,u - |x| - g(|u|)$, which is concave with respect to x. In order to apply Cor. 24.2, then, we need only exhibit a costate p

satisfying (T), (A), and (M) of Theorem 22.26 (with $\eta = 1$). In fact, this would confirm that x_* is a global minimum, since any $\delta > 0$ will serve here.

If $\beta > 0$ (for example), then (T), and (A) yield $p(0) = 0$ and $p'(t) = 1$, whence $p(t) = t$. Then (M) is seen to require that, for every $t \in [0, T]$, we have

$$g(|u|) \geqslant tu \;\forall u \iff u^2 - 2tu + 2 \geqslant 0 \;\forall u \iff t \leqslant \sqrt{2},$$

which is true since $T \leqslant \sqrt{2}$. □

24.4 Exercise. We consider the following problem of L^1 approximation. The goal is to identify the function $x : [0,3] \to \mathbb{R}$ which is nondecreasing and Lipschitz of rank 2, that satisfies $x(0) = 0$, and which best approximates (in the L^1 sense) the function θ given by

$$\theta(t) = \begin{cases} t & \text{if } 0 \leqslant t \leqslant 1 \text{ or } 2 \leqslant t \leqslant 3 \\ 0 & \text{otherwise.} \end{cases}$$

This leads to the following problem of optimal control:

$$\begin{cases} \text{Minimize} & J(x,u) = \displaystyle\int_0^3 |x(t) - \theta(t)|\, dt \\ \text{subject to} & x(0) = 0 \text{ and } x'(t) = u(t) \in [0,2], \; t \in [0,3] \text{ a.e.} \end{cases}$$

Find an optimal process for the problem. □

24.2 Verification functions in control

Consider the familiar optimal control problem

$$\begin{cases} \text{Minimize} & J(x,u) = \ell\big(x(b)\big) + \displaystyle\int_a^b \Lambda\big(t, x(t), u(t)\big)\, dt \\ \text{subject to} & x'(t) = f\big(t, x(t), u(t)\big), \; t \in [a,b] \text{ a.e.} \\ & u(t) \in U, \; t \in [a,b] \text{ a.e.} \\ & x(a) = x_0, \; x(b) \in E. \end{cases} \qquad \textbf{(OC)}$$

Suppose that we wish to confirm the optimality of a certain admissible process (x_*, u_*). The verification function method that we studied in detail in the calculus of variations carries over to the more general setting of (OC), with minor changes. It might be best, if we may so suggest, for the reader to review the discussion in § 19.1 (p. 367) at this point; to a great extent, we content ourselves here with indicating the required modifications.

In the context of (OC), and in its ideal version, the method consists of exhibiting a smooth function φ that satisfies the Hamilton-Jacobi inequality

$$\Lambda(t,x,u)+\varphi_t(t,x)+\langle\,\varphi_x(t,x),f(t,x,u)\rangle \geqslant 0, \quad (t,x,u)\in[a,b]\times\mathbb{R}^n\times U \quad (1)$$

as well as the boundary condition

$$\varphi(b,y) = \ell(y), \quad y\in E. \quad (2)$$

The reader will note the presence of the dynamics function f in (1).

Now let (x,u) be any admissible process. We proceed to express the inequality (1) along (x,u); that is, with $(t,x,u)=(t,x(t),u(t))$, and also with $f(t,x(t),u(t))$ replaced by $x'(t)$ a.e. Then we integrate both sides over $[a,b]$ to obtain

$$\begin{aligned}\int_a^b \Lambda\big(t,x(t),u(t)\big)\,dt \;&\geqslant\; \int_a^b -\frac{d}{dt}\,\varphi(t,x(t))\,dt\\ &= \varphi(a,x(a))-\varphi(b,x(b)) \;=\; \varphi(a,x_0)-\ell(x(b)),\end{aligned}$$

whence $J(x,u)\geqslant\varphi(a,x_0)$. Thus, we have found a lower bound on the cost of any admissible process. If the inequality (1) holds with equality along our suspect process (x_*,u_*), then the lower bound is attained, which verifies that (x_*,u_*) minimizes the cost.

As before, there is a natural candidate for such a function φ, in this case the value function defined by

$$\varphi(\tau,\alpha) \;=\; \min\; \ell\big(x(b)\big)+\int_\tau^b \Lambda\big(t,x(t),u(t)\big)\,dt \;:\; x(\tau)=\alpha\,,\; x(b)\in E. \quad (3)$$

All the difficulties of the verification function method that we had encountered before in § 19.1 (as well as all its advantages) persist in this new setting. In particular, there is a need to consider nonsmooth functions φ. The value function (3) is highly unlikely to be smooth; its behavior at the endpoints of the interval $[a,b]$ is also problematic (again). For these reasons, we introduce a generalized type of solution of the Hamilton-Jacobi inequality for locally Lipschitz verification functions, which turn out, once more, to be highly useful.

The context is that of a locally Lipschitz function $\varphi:\Omega\to\mathbb{R}$, where Ω is an open subset of $[a,b]\times\mathbb{R}^n$. We suppose that φ satisfies the Hamilton-Jacobi inequality in the almost-everywhere sense:

$$\Lambda(t,x,u)+\varphi_t(t,x)+\langle\,\varphi_x(t,x),f(t,x,u)\rangle \geqslant 0 \;\;\forall\,u\in U,\;\;(t,x)\in\Omega \text{ a.e.} \quad (4)$$

The use of such functions is predicated on the following fundamental fact, which allows us to integrate a Hamilton-Jacobi inequality that holds only in this sense, in order to derive a lower bound on the cost.

24.5 Proposition. *We assume that the dynamics function $f(t,x,u)$ is continuous in (t,x), and that the running cost $\Lambda(t,x,u)$ is bounded below on bounded sets,* LB *measurable in (t,u) for each x, and continuous in (t,x) for each u. Let (x,u) be a process for the system (f,U), with x Lipschitz and u essentially bounded, and satisfying $(t,x(t)) \in \Omega \;\; \forall t \in (a,b)$. Then we have*

$$\int_a^b \Lambda(t,x(t),u(t))\,dt \;\geqslant\; \limsup_{\varepsilon \downarrow 0} \left\{ \varphi\big(a+\varepsilon, x(a+\varepsilon)\big) - \varphi\big(b-\varepsilon, x(b-\varepsilon)\big) \right\}. \quad (5)$$

Proof. We first establish an inequality involving the generalized gradient.

Lemma. *For any $(t,x) \in \Omega$ and $u \in U$, we have*

$$\langle (\zeta,\psi), (1, f(t,x,u)) \rangle \leqslant \Lambda(t,x,u) \;\; \forall (\zeta,\psi) \in \partial_C(-\varphi)(t,x).$$

Proof. To see this, recall that (4) holds for the gradient of φ at almost all points in Ω where $-\varphi$ is differentiable, and provides precisely the desired inequality in the case $(\zeta,\psi) = -\nabla\varphi$. But any $(\zeta,\psi) \in \partial_C(-\varphi)(t,x)$ is generated by limiting gradients of $-\varphi$, as described by the gradient formula, Theorem 10.27, a characterization which allows us to ignore any points at which (4) might fail. Since the functions Λ and f are continuous with respect to (t,x), the inequality of the lemma results. □

It follows from the lemma and from the definition of the generalized directional derivative (see Def. 10.3) that we have

$$(-\varphi)^\circ\big(t,x;1,f(t,x,u)\big) = \max\left\{ \zeta + \langle \psi, f(t,x,u) \rangle : (\zeta,\psi) \in -\partial_C\,\varphi(t,x) \right\} \leqslant \Lambda(t,x,u). \quad (6)$$

Now let (x,u) be a process for the system (f,U) as described in the statement of the proposition. When the (locally Lipschitz) function $t \mapsto \varphi(t,x(t))$ is differentiable, when $x'(t)$ exists, and when we have $x'(t) = f(t,x(t),u(t))$ as well as $u(t) \in U$ (thus, for almost every $t \in (a,b)$), we claim that

$$\frac{d}{dt}(-\varphi)\big(t,x(t)\big) \leqslant (-\varphi)^\circ\big(t,x(t);1,f(t,x(t),u(t))\big).$$

It is a simple matter to prove this by examining the difference quotient whose limit is the left side (exercise). Then, integrating the inequality from $a+\varepsilon$ to $b-\varepsilon$, and invoking (6), we derive

$$\int_{a+\varepsilon}^{b-\varepsilon} \frac{d}{dt}(-\varphi)\big(t,x(t)\big)\,dt \leqslant \int_{a+\varepsilon}^{b-\varepsilon} \Lambda\big(t,x(t),u(t)\big)\,dt.$$

Note that the integral of the running cost is well defined, since (by hypothesis) the function $t \mapsto \Lambda(t,x(t),u(t))$ is bounded below, and since it is measurable by Prop. 6.36. Taking the limit as $\varepsilon \downarrow 0$ leads to the inequality (5). □

Remark. Prop. 24.5 extends Prop. 19.2 to the control setting, but with a change of sign, adopted here because, in optimal control, it often seems more natural to vary the initial point rather than the final one. Time reversal allows us to derive one type of result from the other.

In Lipschitz terms, the method of verification functions applies much as it did in the ideal case: the Hamilton-Jacobi inequality holds in a weaker (almost everywhere) sense, but Prop. 24.5 leads to a lower bound on the cost, one which (we hope) is attained by the process (x_*, u_*) that we suspect of being optimal. As we have seen, verification functions can be applied to numerous variants of the basic problem, notably those which feature a unilateral state constraint. In the following, we illustrate the method in the setting of an *infinite horizon* problem.

24.6 Example. Consider the following optimal control problem:

$$\begin{cases} \text{Minimize} & J(x,u) = \displaystyle\int_0^\infty e^{-2t} u(t)(1-x(t))\,dt \\ \text{subject to} & x'(t) = x(t)(4-x(t)) - x(t)\,u(t)\,, \ t \geqslant 0 \text{ a.e.} \\ & u(t) \in [0,4]\,, \ t \geqslant 0 \text{ a.e.} \\ & x(0) = x_0\,. \end{cases}$$

We suppose that the prescribed initial value x_0 lies in $(0,4)$.

Notice that the deductive method cannot be used at our current level of theory, since we have proved in this text neither necessary conditions, nor existence theorems, that apply on the interval $[0,\infty)$. There is no law against formulating a guess, however, as to what the solution might be.

The Hamiltonian is given here (in the normal case) by

$$H(t,x,p,u) = p\left\{x(4-x) - ux\right\} - e^{-2t}u(1-x),$$

which is affine in the control. Let us take the coefficient of u as a switching function:

$$\sigma = e^{-2t}(x-1) - px\,.$$

Assuming without cause that a solution (x,u) exists, and applying the maximum principle without justification, we are led to a costate p for which

$$-p' = p(4-2x-u) + ue^{-2t}, \quad u(t) = \begin{cases} 0 & \text{if } \sigma < 0 \\ 4 & \text{if } \sigma > 0. \end{cases}$$

We ask the reader to be an accomplice in this dubious analysis, by showing that intervals on which σ vanishes correspond to a special value of x:

Exercise. Suppose that $\sigma \equiv 0$ on an interval $[c,d]$. Prove that we have $x \equiv 2$ and $u(t) = 2$ a.e. on $[c,d]$.

It appears that the control values 0, 2, and 4 are destined to play a role in the solution. Based on these clues, and in view of our experience, we are led to formulate an educated conjecture of the turnpike kind: the optimal state arc x is the one which attains the value $x = 2$ as rapidly as possible (by taking $u = 0$ if $x_0 < 2$, or else $u = 4$ if $x_0 > 2$), and then remains at that value thereafter (with $u = 2$).

We now seek to verify this conjecture, by looking at the value function (3) (with $b = +\infty$, and calculated provisionally by means of the conjecture) to see if it has the properties of a verification function. For this purpose, let us make V more explicit.

For $\tau \in \mathbb{R}_+$ and $\alpha \in (0,2]$, we denote by $r(\tau,\alpha)$ the time t at which the solution x of the initial value problem

$$x' = x(4-x),\ x(\tau) = \alpha$$

satisfies $x(t) = 2$. (Note that the solution x converges to the equilibrium value 4, so r is well defined.) Similarly, for $\tau \in \mathbb{R}_+$ and $\alpha \in [2,4)$, we denote by $s(\tau,\alpha)$ the time t at which the solution x of the initial value problem

$$x' = x(4-x) - 4x = -x^2,\ x(\tau) = \alpha$$

satisfies $x(t) = 2$. Then the conjectured optimal cost from an initial condition $(\tau,\alpha) \in \mathbb{R}_+ \times (0,2)$ is given by

$$V^-(\tau,\alpha) = \int_\tau^{r(\tau,\alpha)} 0\,dt + \int_{r(\tau,\alpha)}^\infty e^{-2t}2(1-2)\,dt = -e^{-2r(\tau,\alpha)},$$

whereas the cost from an initial condition $(\tau,\alpha) \in \mathbb{R}_+ \times (2,4)$ is given by

$$V^+(\tau,\alpha) = \int_\tau^{s(\tau,\alpha)} 4e^{-2t}\big(1-x(t)\big)\,dt - e^{-2s(\tau,\alpha)},$$

where, in the integral, x refers to the solution of $x' = -x^2$, $x(\tau) = \alpha$.

The next step is to establish the Hamilton-Jacobi inequality.

Claim 1. For all $(\tau,\alpha) \in \Omega^- := (0,\infty) \times (0,2)$, we have, for all $u \in [0,4]$,

$$e^{-2\tau}u(1-\alpha) + V_\tau^-(\tau,\alpha) + V_\alpha^-(\tau,\alpha)\big\{\alpha(4-\alpha) - u\alpha\big\} \geqslant 0.$$

Note that the function $r(\tau,\alpha)$ is differentiable in Ω^-, by classical results, so the partial derivatives of V^- exist. The inequality of the claim may be written

$$2e^{-2r}\{r_\tau + r_\alpha\,\alpha(4-\alpha)\} + u\big\{e^{-2\tau}(1-\alpha) - 2\alpha e^{-2r}r_\alpha\big\} \geqslant 0. \tag{7}$$

It is clear from the definition of r that $r(\tau,\alpha) = r(0,\alpha) + \tau$, so we have $r_\tau = 1$. Along the solution x of $x' = x(4-x)$, $x(\tau) = \alpha$, we have $r(t,x(t)) = r(\tau,\alpha)$. Differentiating, we find

$$r_\tau\big(t,x(t)\big)+r_\alpha\big(t,x(t)\big)\,x(t)\big(4-x(t)\big) = 0.$$

Substituting $t=\tau$ yields $r_\alpha(\tau,\alpha) = -\big\{\alpha(4-\alpha)\big\}^{-1}$. It follows that the first expression in (7) vanishes, so that (7) is equivalent to

$$e^{-2\tau}(1-\alpha)-2\,\alpha\,e^{-2r}r_\alpha = e^{-2\tau}\{1-\alpha+2/(4-\alpha)\} \geqslant 0,$$

an inequality which is easily seen to hold. This proves Claim 1.

Claim 2. For all $(\tau,\alpha) \in \Omega^+ := (0,\infty)\times(2,4)$, we have, for all $u\in[0,4]$,

$$e^{-2\tau}u(1-\alpha)+V^+_\tau(\tau,\alpha)+V^+_\alpha(\tau,\alpha)\big\{\,\alpha(4-\alpha)-u\,\alpha\,\big\} \geqslant 0.$$

It follows, much as in the previous case, using now the identities

$$s_\alpha(\tau,\alpha) = 1/\alpha^2, \quad s(\tau,\alpha) = s(0,\alpha)+\tau,$$

that the claim reduces to proving the inequality

$$\big[1+(2/\alpha)\,e^{-2s(0,\alpha)}-\alpha\big]\,u \geqslant 4(1-\alpha)+8\,e^{-2s(0,\alpha)}/\alpha.$$

The coefficient of u on the left is negative, so we may set $u=4$ in proving the last inequality. But in that case it reduces to an identity; whence Claim 2.

We now define, for $(\tau,\alpha) \in \Omega := (0,\infty)\times(0,4)$, the function

$$V(\tau,\alpha) = \begin{cases} V^-(\tau,\alpha) & \text{if } (\tau,\alpha)\in\Omega^- \\ -e^{-2\tau} & \text{if } \alpha = 2 \\ V^+(\tau,\alpha) & \text{if } (\tau,\alpha)\in\Omega^+. \end{cases}$$

Then V is locally Lipschitz (see Exer. 13.18) and satisfies the Hamilton-Jacobi inequality (4) in the almost-everywhere sense in Ω. Since any admissible process (x,u) with $x(0)\in(0,4)$ is such that $(t,x(t))\in\Omega\ \ \forall t>0$, we deduce with the help of Prop. 24.5 that for any $T>0$ we have

$$\int_0^T e^{-2t}u(t)\big(1-x(t)\big)\,dt \geqslant \limsup_{\varepsilon\downarrow 0}\big\{V\big(\varepsilon,x(\varepsilon)\big)-V\big(T-\varepsilon,x(T-\varepsilon)\big)\big\}.$$

It is not difficult to verify that we have, uniformly for $\alpha\in(0,4)$,

$$V(\tau,\alpha)\to V(0,\alpha) \text{ as } \tau\downarrow 0, \quad V(\tau,\alpha)\to 0 \text{ as } \tau\uparrow+\infty.$$

This allows us to deduce, by letting $T\to\infty$ in the preceding inequality,

$$\int_0^\infty e^{-2t}u(t)\big(1-x(t)\big)\,dt \geqslant V(0,x_0).$$

(The integral is well defined, since u and x are bounded.) Since this holds for any admissible process (x,u), and since our conjecture yields the cost $V(0,x_0)$ by construction, we have verified the validity of our guess. □

24.3 Use of the Hamilton-Jacobi equation

We shall illustrate in this section how a uniqueness theorem for generalized solutions of the Hamilton-Jacobi equation leads to an inductive method to confirm conjectured optimality. We consider an autonomous, finitely generated control system

$$f(x,u) = g_0(x) + G(x)\,u = g_0(x) + \sum_{j=1}^{m} g_j(x)\,u^j, \ \ u(t) \in U,$$

where $U \subset \mathbb{R}^m$ is compact and convex, and where the functions g_j have linear growth. The lower Hamiltonian h of the system is the function

$$h(x,p) = \langle p, g_0(x)\rangle + \min\big\{ \langle p, G(x)\,u\rangle : u \in U \big\}.$$

The *minimal-time function* $T : \mathbb{R}^n \to \mathbb{R}_\infty$ is defined as follows: $T(\alpha)$ is the infimum of all $\tau \geqslant 0$ such that, for some trajectory x of the system, we have

$$x(0) = \alpha\,,\ x(\tau) = 0.$$

The reader will recall that a real-valued function φ is said to be *positive definite* when $\varphi(0) = 0$ and $\varphi(x) > 0$ for $x \neq 0$.

24.7 Theorem. *Let φ be a continuous, positive definite function that satisfies*

$$x \neq 0,\ \zeta \in \partial_P\,\varphi(x) \implies h(x,\zeta) = -1\,.$$

Then φ is the minimal-time function.

Proof. We consider the augmented state $x_+ = (x^0, x) \in \mathbb{R}\times\mathbb{R}^n$ in which the new coordinate x^0 is a surrogate for time t. We define a multifunction F_+ and a function φ_+ as follows:

$$F_+(x^0,x) = \big\{\big(1, g_0(x) + G(x)\,u\big) : u \in U\big\},\ \ \varphi_+\big(x^0,x\big) = x^0 + \varphi(x)\,.$$

Then the lower Hamiltonian of F_+ is the function $h_+(x_+, p_+) = p^0 + h(x,p)$, and the hypothesis implies

$$h_{F_+}\big(x_+, \partial_P\,\varphi_+(x_+)\big) = 0 \ \ \forall\, x_+ \in \Omega := \{(x^0,x) : x^0 \in \mathbb{R},\ x \in \mathbb{R}^n\backslash\{0\}\}.$$

It follows that the system (φ_+, F_+) is both strongly increasing and weakly decreasing relative to Ω (see Theorems 12.11 and 12.17, and Exer. 13.30).

Fix $\alpha \in \mathbb{R}^n \backslash \{0\}$, and consider any state trajectory x for the original system, with $x(0) = \alpha$, that attains zero in finite time. Let $\tau > 0$ be the first time for which $x(\tau)$ equals 0. Note that $(t, x(t))$ is a trajectory for F_+ on $[0, \tau)$ that lies in Ω. Strong increase implies

$$\tau + \varphi\big(x(\tau)\big) \geqslant 0 + \varphi\big(x(0)\big) \implies \tau \geqslant \varphi(\alpha).$$

It follows that $T(\alpha) \geqslant \varphi(\alpha)$. We need only establish the opposite inequality.

By weak decrease (see Theorem 12.11), there is a trajectory $(t, x(t))$ for F_+ beginning at $(0, \alpha)$, maximally defined for Ω, with the property that

$$t + \varphi\big(x(t)\big) \leqslant \varphi(\alpha),\ t \in [0, \bar{t}],$$

where $\bar{t}$ is the exit time from Ω. Since $\varphi \geqslant 0$, it follows that, for some positive τ no greater than $\varphi(\alpha)$, we have $\varphi(x(\tau)) = 0$; that is, $x(\tau) = 0$ (by positive definiteness). Filippov's lemma assures us that x is a trajectory of the original control system. Thus, $T(\alpha) \leqslant \tau \leqslant \varphi(\alpha)$. □

Remark. Observe that no *a priori* assumption is made in Theorem 24.7 about the controllability of the system to 0; it is the existence of φ that forces T to be finite-valued, or, equivalently, guarantees that every point in $\mathbb{R}^n$ can be steered to 0 in finite time. Note also the need to exclude $x = 0$ in the proximal Hamilton-Jacobi equation: since φ attains a minimum at the origin, we have $0 \in \partial_P \varphi(0)$, but $h(0,0) = 0 \neq -1$. Thus the Hamilton-Jacobi equation necessarily fails at 0.

A converse to Theorem 24.7 can be proved:

24.8 Exercise. Suppose that the minimal-time function T is finite and continuous (for a system (f, U) as above). Show that it is positive definite and satisfies

$$x \neq 0,\ \zeta \in \partial_P T(x) \implies h(x, \zeta) = -1. \qquad \square$$

What bearing does a result such as Theorem 24.7 have on sufficient conditions for optimality? The answer lies in the following observation. If we formulate a conjecture regarding the minimal-time path from any initial value α, and if we calculate the resulting time $T(\alpha)$ based on the conjecture, then the conjecture is correct if and only if the function we calculated coincides with the minimal-time function. And this can be checked by verifying the properties that are known to characterize that function. We illustrate the procedure now.

24.9 Example. Consider again the soft landing problem (Example 22.14). As we saw, the function claimed to be the minimal-time function is given in terms of the switching curve Σ as follows:

$$\varphi(x,y) = \begin{cases} -y+\sqrt{2y^2-4x} & \text{if } (x,y) \text{ is left of } S\text{: } 2x \leqslant -y^2 \text{ and } y \geqslant 0, \text{ or} \\ & \quad 2x \leqslant y^2 \text{ and } y \leqslant 0 \\ +y+\sqrt{2y^2+4x} & \text{if } (x,y) \text{ is right of } S\text{: } 2x \geqslant -y^2 \text{ and } y \geqslant 0, \text{ or} \\ & \quad 2x \geqslant y^2 \text{ and } y \leqslant 0. \end{cases}$$

It was calculated on the basis of information derived from the maximum principle. We may verify the optimality of the proposed strategy by showing that φ is in fact the minimal-time function. We do this by showing that it has the properties given in Theorem 24.7.

It follows easily that φ is continuous and positive definite; it is the proximal Hamilton-Jacobi equation that needs to be verified. Note that

$$h(x,y,p,q) = py-|q|.$$

If (x,y) does not lie on the switching curve Σ, then φ is smooth and $\partial_P\,\varphi(x,y)$ is the singleton gradient, easily calculated. We check without difficulty that

$$h\big(x,y,\varphi_x(x,y),\varphi_y(x,y)\big) = \varphi_x(x,y)\,y-|\,\varphi_y(x,y)| = -1.$$

Consider now a point $(y^2/2\,,y)$, with $y < 0$; this is a point on the lower branch of Σ. Let (p,q) belong to $\partial_P\,\varphi(x,y)$. Then, for some $\sigma \geqslant 0$, the proximal inequality asserts that locally, relative to all points (X,Y) to the right of Σ (that is, satisfying $2Y^2-4X \leqslant 0)$, the function

$$(X,Y) \mapsto Y+\sqrt{2Y^2+4X}-pX-qY+\sigma\big\{\,|X-x|^2+|Y-y|^2\big\}$$

attains a minimum at (x,y). The multiplier rule (Theorem 9.1) yields the existence of $\gamma \geqslant 0$ such that

$$p = -2\gamma-1/y, \quad q = 2\gamma y \leqslant 0,$$

whence

$$h(x,y,p,q) = py-|q| = py+q = -1,$$

as required. A similar argument applies to the upper branch of Σ. It follows then, from Theorem 24.7, that φ *is* the minimal-time function, and thus that the strategy generating it is optimal. □

The approach we have described in this section can be used in any context for which we happen to have the appropriate characterization of the value function.

Chapter 25
Differential inclusions

Proof of the Multiplier Rule is both formidable and tedious, and we shall sketch only a part of it here. After completing this chapter, a reader will have acquired a feeling for what the rule says and does and it will then be easier to endure the details of a complete proof. Those who have a serious interest in variational theory must, sooner or later, study some of the proofs.

G. M. Ewing (Calculus of Variations with Applications)

We develop in this chapter certain necessary conditions for the optimal control of systems that are described by a differential inclusion. The discussion takes place in the context of the following deceptively simple-looking problem:

$$\left\{\begin{array}{ll} \text{Minimize} & J(x) = \ell\big(x(a), x(b)\big) \\ \text{subject to} & x'(t) \in F_t\big(x(t)\big),\ t \in [a,b] \text{ a.e.} \\ & \big(x(a), x(b)\big) \in E. \end{array}\right. \qquad \textbf{(DI)}$$

Here, F is a multifunction mapping $[a,b]\times\mathbb{R}^n$ to the subsets of $\mathbb{R}^n$. The reader will notice that the t-dependence of F is indicated by a subscript (as it will be for other data subsequently). An arc $x : [a,b] \to \mathbb{R}^n$ is said to be *admissible* for the problem if it satisfies the differential inclusion and the boundary condition of (DI).

We have met differential inclusions before, and have found them useful in studying such topics as invariance, monotonicity, and relaxation. Nonetheless, we feel obliged to admit that the differential inclusion problem (DI) is less natural than the standard control formulation (for one thing, the control variable has disappeared), and is rarely used in the modeling of applications. We must beg to be trusted in a matter such as this, however: (DI) provides an ideal mathematical environment in which to prove various types of necessary conditions.

All the different versions of the maximum principle presented so far have followed from the extended maximum principle; this, in turn, will be a consequence of the first theorem proved below. In later sections, we shall derive advanced multiplier rules in optimal control.

Uniquely, this chapter contains no exercises. The author feels that the reader will be sufficiently exercised in working through the proofs, which are the most technical ones in the book.

F. Clarke, *Functional Analysis, Calculus of Variations and Optimal Control*,
Graduate Texts in Mathematics 264, DOI 10.1007/978-1-4471-4820-3_25,

25.1 A theorem for Lipschitz multifunctions

An arc x_* which is admissible for the problem (DI) is said to be a *local minimizer* if, for some $\varepsilon > 0$, we have $J(x_*) \leqslant J(x)$ whenever x is an admissible arc satisfying $\|x - x_*\| \leqslant \varepsilon$. (As usual, the supremum or L^∞ norm is meant here.)

Notation: For each $t \in [a,b]$, G_t denotes the graph of the multifunction $F_t(\cdot)$:

$$G_t = \{(x,v) \in \mathbb{R}^n \times \mathbb{R}^n : v \in F_t(x)\}.$$

We posit the following hypotheses relative to a given local minimizer x_* of (DI):

[H1] The function ℓ is locally Lipschitz; the set E is closed; the multifunction $t \mapsto G_t$ is measurable; for some $\delta > 0$, the following set is closed for almost every t:

$$\{(x,v) \in G_t : |x - x_*(t)| \leqslant \delta\}.$$

[H2] There exists a summable function k such that, for almost every t,

$$x,y \in B(x_*(t),\delta) \Longrightarrow F_t(y) \subset F_t(x) + B(0, k_t|x-y|).$$

The reader will observe that the first hypothesis forces the values of F_t to be closed sets, near $x_*(t)$ at least. The second one asserts that F_t satisfies a Lipschitz condition in the sense of multifunctions (see Def. 12.14).

25.1 Theorem. *Let x_* be a local minimizer for the problem* (DI), *under hypotheses* [H1][H2] *above. Then there exist an arc p and a scalar η equal to* 0 *or* 1 *satisfying the* **nontriviality condition**

$$(\eta, p(t)) \neq 0 \;\; \forall t \in [a,b], \tag{25.1 a}$$

the **transversality condition**

$$(p(a), -p(b)) \in \eta\, \partial_L \ell(x_*(a), x_*(b)) + N_E^L(x_*(a), x_*(b)), \tag{25.1 b}$$

the **Euler inclusion** *for almost every t*:

$$p'(t) \in \operatorname{co}\{\omega : (\omega, p(t)) \in N_{G_t}^L(x_*(t), x_*'(t))\} \text{ a.e.} \tag{25.1 c}$$

as well as the **maximum condition** *for almost every t*:

$$\langle p(t), v\rangle \leqslant \langle p(t), x_*'(t)\rangle \;\; \forall v \in F_t(x_*(t)). \tag{25.1 d}$$

We remark that the Euler inclusion may be recognized as having the same form as that of Theorem 18.13, if one recalls that $\partial_L I_{G_t} = N_{G_t}^L$, where I_{G_t} is the indicator function of G_t.

The following technical result on Lipschitz multifunctions will be needed in the proof of Theorem 25.1, and later on as well.

25.2 Proposition. *Let Γ be a multifunction from $\mathbb{R}^n$ to $\mathbb{R}^m$ with closed graph G, and let d_G denote the Euclidean distance function of G. Suppose that Γ satisfies the local Lipschitz condition*

$$\Gamma(y) \subset \Gamma(z) + B\big(0, k|y-z|\big) \;\; \forall\, y,z \in B(x_0,r), \tag{1}$$

where $x_0 \in \mathbb{R}^n$, $r > 0$. Let $v_0 \in \Gamma(x_0)$. Then

$$(\alpha,\beta) \in N_G^L\big(x_0,v_0\big) \implies |\alpha| \leqslant k|\beta|. \tag{2}$$

If the Lipschitz condition in (1) holds globally (that is, for all $y,z \in \mathbb{R}^n$), then for any point $(x,v) \in \mathbb{R}^n \times \mathbb{R}^m$, we have

$$(\alpha,\beta) \in \partial_L d_G(x,v) \implies |\alpha| \leqslant k|\beta|, \tag{3}$$

and

$$d_G(x,v) > 0,\, (\alpha,\beta) \in \partial_L d_G(x,v) \implies |\beta| \geqslant (1+k^2)^{-1/2}. \tag{4}$$

Proof. As regards (2), it suffices to consider $(\alpha,\beta) \in N_G^P(x_0,v_0)$, since N_G^P generates N_G^L via limits. In turn, it suffices to consider points (α,β) belonging to $\partial_P d_G(x_0,v_0)$, since $N_G^P(x_0,v_0)$ is the cone generated by this set (Prop. 11.28). In that case, the proximal subgradient inequality asserts that for some $\sigma \geqslant 0$, for all (x,v) sufficiently near (x_0,v_0), we have

$$d_G(x,v) + \sigma|(x-x_0, v-v_0)|^2 \geqslant \langle (\alpha,\beta), (x-x_0, v-v_0)\rangle.$$

For any x near x_0, there exists $v \in \Gamma(x)$ such that $|v-v_0| \leqslant k|x-x_0|$, by hypothesis (1). For all x sufficiently close to x_0, this choice of (x,v) may be substituted in the proximal inequality above. Since $d_G(x,v) = 0$, doing so leads to

$$\langle \alpha, x-x_0\rangle \leqslant \langle \beta, v_0-v\rangle + \sigma\big(|x-x_0|^2 + |v-v_0|^2\big) \leqslant \\ |\beta|k|x-x_0| + \sigma\big(1+k^2\big)|x-x_0|^2$$

for all x in a neighborhood of x_0. This yields $|\alpha| \leqslant k|\beta|$, and confirms (2).

We turn now to (3) and (4), for which only the case $d_G(x,v) > 0$ need be considered. Then (α,β) is of the form

$$(\alpha,\beta) = (x-\bar{x}, v-\bar{v})/|(x-\bar{x}, v-\bar{v})|,$$

where $(\bar{x},\bar{v})$ is a closest point in G to (x,v) (see Exer. 13.19). By Prop. 11.34 we have $(\alpha,\beta) \in N_G^L(\bar{x},\bar{v})$, whence (by (2)) we deduce $|\alpha| \leqslant k|\beta|$, yielding (3). But we also have, as a consequence of the characterization above, $|\alpha|^2 + |\beta|^2 = 1$. Combined with $|\alpha| \leqslant k|\beta|$, this implies the lower bound on $|\beta|$ stated in (4). □

The proposition implies that the set whose convex hull is taken in the Euler inclusion (25.1 c) is compact, which is why we do not require the *closed* convex hull later: the convex hull is already closed (by Exer. 2.8).

The following sequential closure result will be used in the proof, and again later.

25.3 Proposition. *Let $t \mapsto G_t$ be a measurable closed-valued multifunction from $[a,b]$ to $\mathbb{R}^n \times \mathbb{R}^n$ that satisfies the bounded slope condition*[1] *(for almost every t)*

$$(x,v) \in G_t\,,\ (\alpha,\beta) \in N^L_{G_t}\big(x,v\big) \implies |\,\alpha\,| \leqslant k(t)|\,\beta\,|,$$

where k is summable. Let x_i, v_i, p_i, and q_i be measurable functions such that

$$q_i \to q \ \text{ weakly in } L^1(a,b),\ \ p_i(t) \to p(t) \text{ a.e.},\ \ x_i(t) \to x(t) \text{ a.e.},\ \ v_i(t) \to v(t) \text{ a.e.}$$

and satisfying, for some constant M, for each i:

$$|\,p_i(t)| \leqslant M \text{ a.e.},\ \ |\,q_i(t)| \leqslant k(t) \text{ a.e.}$$

Suppose that for each i we have $(x_i(t),v_i(t)) \in G_t$ a.e. and

$$q_i(t) \in \operatorname{co}\big\{\,\omega : \big(\omega, p_i(t)\big) \in N^L_{G_t}\big(x_i(t),v_i(t)\big) + \varepsilon_i B\,\big\},\ \ t \in \Omega_i \text{ a.e.}, \tag{5}$$

where $\varepsilon_i \downarrow 0$, and where Ω_i is a sequence of measurable subsets of $[a,b]$ such that $\operatorname{meas}(\Omega_i) \to b-a$. *Then we have in the limit*

$$q(t) \in \operatorname{co}\big\{\,\omega : \big(\omega, p(t)\big) \in N^L_{G_t}\big(x(t),v(t)\big)\big\} \text{ a.e.}$$

Proof. Fix i sufficiently large so that $\varepsilon_i < 1$. For almost every t, any point $(\omega, p_i(t))$ as described in (5) satisfies

$$\big(\omega - u\,, p_i(t) - w\big) \in N^L_{G_t}\big(x_i(t),v_i(t)\big)$$

for some point $(u\,,w) \in \varepsilon_i B$. The bounded slope condition implies

$$|\,\omega - u\,| \leqslant k(t)|\,p_i(t) - w\,|,$$

whence

$$\big|(\omega - u\,, p_i(t) - w)\big| \leqslant (k(t)+1)\,(M+1) =: R(t).$$

Then, by Prop. 11.34, we have

$$\begin{aligned}\big(\omega - u\,, p_i(t) - w\big) &\in \big|\big(\omega - u\,, p_i(t) - w\big)\big|\,\partial_L d_{G_t}\big(x_i(t),v_i(t)\big)\\ &\subset [\,0\,,R(t)]\,\partial_L d_{G_t}\big(x_i(t),v_i(t)\big).\end{aligned}$$

This implies in turn

[1] This is not to be confused with the very different "bounded slope condition" of Def. 20.17.

$$\big(\omega, p_i(t)\big) \in [0,1]\, R(t)\, \partial_L d_{G_t}\big(x_i(t), v_i(t)\big) + \varepsilon_i B .$$

Thus, we have for all i sufficiently large,

$$q_i(t) \in \operatorname{co}\big\{ \omega : \big(\omega, p_i(t)\big) \in [0,1]\, R(t)\, \partial_L d_{G_t}\big(x_i(t), v_i(t)\big) + \varepsilon_i B \big\},\ t \in \Omega_i \text{ a.e.}$$

We now seek to invoke Prop. 18.6, with $f(t,x,v) = R(t)\, d_{G_t}(u,v)$. The required measurability hypothesis follows from Exer. 6.30, and the Lipschitz hypothesis holds because a distance function is globally Lipschitz of rank 1. We deduce that in the limit, we have almost everywhere

$$\begin{aligned} q(t) &\in \operatorname{co}\big\{ \omega : \big(\omega, p(t)\big) \in [0,1]\, R(t)\, \partial_L d_{G_t}\big(x(t), v(t)\big)\big\} \\ &\subset \operatorname{co}\big\{ \omega : \big(\omega, p(t)\big) \in N^L_{G_t}\big(x(t), v(t)\big)\big\}. \end{aligned}$$

The last inclusion follows from Prop. 11.34, since $(x(t), v(t)) \in G_t$ a.e., as a result of the fact that G_t is closed-valued. □

A reduction. It turns out that it suffices to prove Theorem 25.1 for the case in which the δ of hypotheses [H1] and [H2] (originally taken to be finite) is $+\infty$. We see this by considering the multifunction

$$\widetilde{F}_t(x) = F_t(\pi_t(x)),$$

where $\pi_t(x)$ denotes the projection of x onto the set $B(x_*(t), \delta)$. The mapping $(t,x) \mapsto \pi_t(x)$ is measurable in t and globally Lipschitz of rank 1 in x (see p. 353). Note that F_t and $\widetilde{F}_t$ agree on $B(x_*(t), \delta)$, so x_* continues to be a local minimizer for the version of (DI) in which F is replaced by $\widetilde{F}$. But this modified multifunction satisfies [H1] and [H2] with $\delta = +\infty$. Furthermore, the conclusions of the theorem for $\widetilde{F}$ coincide with those for F, since the graphs of F_t and $\widetilde{F}_t$ agree locally near $(x_*(t), x_*'(t))$. For these reasons, we may (and do) make the useful assumption from now on that [H1] and [H2] hold with $\delta = +\infty$.

Proof of Theorem 25.1. We proceed to prove the theorem in the presence of two additional, temporary, hypotheses. The removal of these temporary hypotheses will be the final step in the proof.

[TH1] k is a constant function.

[TH2] $\ell(x_1, x_2)$ is of the form $\ell(x_2)$, and E is of the form $C_0 \times C_1$.

Let ε_i be a positive sequence decreasing to 0. For fixed i, let us consider the problem P_i of minimizing

$$J_i(x) = \ell_i\big(x(b)\big) + (1/\varepsilon_i) \int_a^b d_{G_t}\big(x(t), x'(t)\big)\, dt$$

over the set S of arcs x satisfying

$$x(a) \in C_0\,,\ x(b) \in C_1\,,\ \|x - x_*\| \leqslant \varepsilon\,,$$

where ℓ_i is defined by

$$\ell_i(x) = \big[\,\ell(x) - \ell\big(x_*(b)\big) + \varepsilon_i^2\,\big]_+ .$$

(Notation: $[\,a\,]_+ = \max\{0, a\}$.) Note that [TH2] is used in defining the problem P_i, and that the measurability of G_t implies that the integral is well defined.

Because C_0 and C_1 are closed (since E is closed by hypothesis), the set S is a complete metric space when equipped with the metric

$$d(x,y) = \big|x(a) - y(a)\big| + \|x' - y'\|_1 = \big|x(a) - y(a)\big| + \int_a^b |x'(t) - y'(t)|\,dt\,.$$

It is clear that the infimum in the problem P_i is necessarily nonnegative, and that x_* assigns the value ε_i^2 to the cost J_i (which is lower semicontinuous). Thus x_* yields the infimum within ε_i^2.

By Theorem 5.19, there exists an arc $x_i \in S$ satisfying

$$\big|x_i(a) - x_*(a)\big| + \|x_i' - x_*'\|_1 \leqslant \varepsilon_i$$

and such that x_i minimizes over S the perturbed cost

$$\ell_i(x(b)) + \varepsilon_i\big|x(a) - x_i(a)\big| + (1/\varepsilon_i)\int_a^b d_{G_t}\big(x(t), x'(t)\big)\,dt + \varepsilon_i\int_a^b \big|x'(t) - x_i'(t)\big|\,dt\,.$$

We have $\|x_i - x_*\| + \|x_i' - x_*'\|_1 < \varepsilon$ for i sufficiently large, and it follows that x_i is a local minimum in the sense of Theorem 18.1 for the perturbed problem above; it is clear that the theorem applies (with $V_t \equiv \mathbb{R}^n$). We deduce the existence of an arc p_i such that

$$-p_i(b) \in \partial_L\,\ell_i\big(x_i(b)\big) + N_{C_1}^L\big(x_i(b)\big),\ \ p_i(a) \in N_{C_0}^L(x_i(a)) + \varepsilon_i B \tag{6}$$

$$p_i'(t) \in \mathrm{co}\,\big\{\,\omega : \big(\omega, p_i(t)\big) \in (1/\varepsilon_i)\,\partial_L\,d_{G_t}\big(x_i(t), x_i'(t)\big) + \{0\}\times\varepsilon_i B\,\big\} \text{ a.e.} \tag{7}$$

$$\begin{aligned}\langle\, p_i(t), v\,\rangle - (1/\varepsilon_i)\,d_{G_t}\big(x_i(t), v\big) - \varepsilon_i\big|v - x_i'(t)\big| \\ \leqslant \langle\, p_i(t), x_i'(t)\,\rangle - (1/\varepsilon_i)\,d_{G_t}\big(x_i(t), x_i'(t)\big)\ \ \forall\, v \text{ a.e.}\end{aligned} \tag{8}$$

Invoking Prop. 25.2 together with [H2] (which holds with $\delta = +\infty$, let us recall) and the Euler inclusion (7) reveals

$$\big|\,p_i'(t)\big| \leqslant k\big(\,|\,p_i(t)| + \varepsilon_i\,\big) \text{ a.e.} \tag{9}$$

Note also that (8) implies, for almost every t:

$$\langle\, p_i(t), v\,\rangle - \varepsilon_i\big|v - x_i'(t)\big| \leqslant \langle\, p_i(t), x_i'(t)\,\rangle\ \ \forall\, v \in F_t\big(x_i(t)\big)\,. \tag{10}$$

Convergence. By taking subsequences as necessary (without relabeling), we may arrange that either

$$\int_a^b d_{G_t}(x_i, x_i')\,dt \; > 0 \;\; \forall\, i,$$

or else that the integral is zero for every i. We also arrange to have x_i' converge almost everywhere to x_*'.

Case 1: $\displaystyle\int_a^b d_{G_t}\big(x_i, x_i'\big)\,dt > 0 \;\; \forall\, i.$

In this case, there is for each i a set S_i of positive measure on which $d_{G_t}(x_i, x_i') > 0$. By Prop. 25.2, and in light of (7), we deduce

$$\frac{1/\varepsilon_i}{\sqrt{1+k^2}} - \varepsilon_i \leqslant |\,p_i(t)| \leqslant 1/\varepsilon_i + \varepsilon_i\,, \;\; t \in S_i \text{ a.e.} \tag{11}$$

We proceed to rewrite (6) (7) (8) with p_i replaced by $\varepsilon_i\, p_i$ (that is, we multiply across by ε_i, without relabeling):

$$-p_i(b) \in \varepsilon_i\, \partial_L \ell_i\big(x_i(b)\big) + N^L_{C_1}\big(x_i(b)\big), \;\; p_i(a) \in N^L_{C_0}(x_i(a)) + \varepsilon_i^2 B \tag{12}$$

$$p_i'(t) \in \operatorname{co}\big\{\, \omega : \big(\omega, p_i(t)\big) \in \partial_L d_{G_t}\big(x_i(t), x_i'(t)\big) + \{0\}\times \varepsilon_i^2 B \,\big\} \text{ a.e.} \tag{13}$$

$$\begin{aligned} &\langle\, p_i(t), v\rangle - d_{G_t}\big(x_i(t), v\big) - \varepsilon_i^2\big|v - x_i'(t)\big| \\ &\quad \leqslant \langle\, p_i(t), x_i'(t)\rangle - d_{G_t}\big(x_i(t), x_i'(t)\big) \;\; \forall\, v \text{ a.e.} \end{aligned} \tag{14}$$

The inequalities (9) and (11) transform as follows:

$$\big|\,p_i'(t)\big| \leqslant k\big(\,|\,p_i(t)| + \varepsilon_i^2\,\big) \text{ a.e.} \tag{15}$$

$$\frac{1}{\sqrt{1+k^2}} - \varepsilon_i^2 \leqslant |\,p_i(t)| \leqslant 1 + \varepsilon_i^2\,, \;\; t \in S_i \text{ a.e.} \tag{16}$$

These two facts allow us to deduce that (for a subsequence), p_i converges uniformly to an arc p and that p_i' converges weakly in L^1 to p', essentially as in Exer. 6.42. Note that p is nonzero; in fact, we have $\|\,p\,\| \geqslant (1+k^2)^{-1/2}$ as a consequence of the first bound in (16).

The passage to the limit in (13) is justified by invoking Prop. 18.6, where we take $f(t,x,v) = d_{G_t}(x,v)$. Note that G_t is closed for almost every t as a consequence of [H1], and that the mapping $t \mapsto d_{G_t}(x,v)$ is measurable by Exer. 6.30. It follows that the arc p satisfies

$$p'(t) \in \operatorname{co}\big\{\, \omega : \big(\omega, p(t)\big) \in \partial_L d_{G_t}\big(x_*(t), x_*'(t)\big) \big\} \text{ a.e.} \tag{17}$$

This implies (25.1 c), since, when $(x,v) \in G_t$, the cone $N^L_{G_t}(x,v)$ is the one generated by $\partial_L d_{G_t}(x,v)$; see Prop. 11.34.

A further consequence of the analysis is that p satisfies $|p'(t)| \leqslant k|p(t)|$ a.e., as a result of (15). This implies that p is nonvanishing, or else p would be identically zero by Gronwall's lemma, contradicting $\|p\| \geqslant (1+k^2)^{-1/2}$. This yields (25.1 a). Finally, it is clear that (12) leads to (25.1 b), with $\eta = 0$ (since N^L has closed graph), and that in the limit, (14) gives rise to (25.1 d). All the conclusions of the theorem are verified.

Case 2: $\displaystyle\int_a^b d_{G_t}\big(x_i, x_i'\big)\,dt = 0 \;\; \forall\, i.$

It follows in this case that x_i is a trajectory for F. Then $\ell_i(x_i(b)) > 0 \;\; \forall\, i$, for otherwise the optimality of x_* is contradicted. Consequently, we have

$$\ell_i(x) = \ell(x) - \ell\big(x_*(b)\big) + \varepsilon_i^2$$

for x in a neighborhood of $x_i(b)$, so that

$$\partial_L \ell_i(x_i(b)) = \partial_L \ell(x_i(b)). \tag{18}$$

Observe that in this Case 2, (7) implies (by Prop. 11.34)

$$p_i'(t) \in \operatorname{co}\big\{\, \omega : \big(\omega, p_i(t)\big) \in N^L_{G_t}\big(x_i(t), x_i'(t)\big) + \{0\} \times \varepsilon_i B \,\big\} \text{ a.e.} \tag{19}$$

We may identify two subcases (by taking further subsequences):

$$\|p_i\| \text{ is bounded}; \quad \|p_i\| \to \infty.$$

In the *first subcase*, Gronwall's lemma together with (9) allows us to deduce again that (for a subsequence), p_i converges uniformly to an arc p and p_i' converges weakly in L^1 to p'. We invoke Prop. 25.3 to pass to the limit in (19), and we deduce that the arc p satisfies (25.1 c). In view of (18), it is clear that (6) leads to (25.1 b), with $\eta = 1$; consequently, (25.1 a) holds. There remains (25.1 d) to confirm.

Fix any t for which $x_i'(t) \to x_*'(t)$ as well as $x_i'(t) \in F_t(x_i(t)) \;\; \forall\, i$, for which [H2] holds, and for which (10) holds for all i (these conditions hold on a set of full measure). Now choose any $v \in F_t(x_*(t))$. For each i, by the Lipschitz property of F_t, there exists $v_i \in F_t(x_i(t))$ such that

$$|v_i - x_i'(t)| \leqslant k|x_i(t) - x_*(t)|.$$

Then the inequality (10) holds for this value v_i. Passing to the limit, we deduce that $\langle p(t), v\rangle \leqslant \langle p(t), x_*'(t)\rangle$, as required.

In the *second subcase*, when $\|p_i\| \to \infty$, we divide by $\|p_i\|$ in (6) (10) (19); that is, we replace p_i throughout by $p_i/\|p_i\|$. Then the same convergence argument as above applies, giving rise to a nonvanishing limiting arc p with $\|p\| = 1$ satisfying the required necessary conditions (but now with $\eta = 0$). It follows as it did at the end of Case 1 that p is nonvanishing.

Removal of the temporary hypotheses. We have proved Theorem 25.1 under the temporary hypotheses [TH1] [TH2]. Suppose now that the problem satisfies [TH1] but not [TH2].

We introduce a reformulation device in which the extended state variable is (x,y). The role of y will be as stand-in for $x(a)$. The new multifunction F^+ and cost function ℓ^+ are given by

$$F_t^+(x,y) = \{ (v,0) : v \in F_t(x) \}, \;\; \ell^+(x,y) = \ell(y,x),$$

and the boundary constraints are specified by

$$C_0^+ = \{ (x,y) : x = y \}, \;\; C_1^+ = \{ (x,y) : (y,x) \in E \}.$$

It is an easy matter to check that the extended arc $(x_*, x_*(a))$ is a local minimizer for the corresponding problem. The hypotheses [H1] [H2] are present, together with *both* [TH1] and [TH2]. One may therefore apply the theorem to the extended problem; routine analysis of the resulting necessary conditions leads to the desired conclusions.

We now remove the sole remaining temporary hypothesis [TH1], by showing that it suffices to prove the theorem in the case in which the function k, which may always be assumed to be strictly positive, is identically 1. Without loss of generality, to ease the notation, we suppose that $x_* \equiv 0$. (The reduction to this case replaces $F_t(x)$ by $F_t(x_*(t)+x) - x_*'(t)$, with the evident translations applied to ℓ and E.)

We shall use the change of time scale induced by $s = \tau(t)$, where

$$\tau(t) := \int_a^t k(\sigma)\,d\sigma,$$

and we define the data of a rescaled version of the problem by setting

$$\widetilde{F}_s(y) = \frac{1}{k_t} F_t(y), \text{ where } t = \tau^{-1}(s).$$

The transformed problem is to be considered relative to arcs y on $[0,T]$, where $T := \tau(b)$; ℓ and E are unchanged.

Lemma. *The arc $y_* \equiv 0$ solves the transformed problem relative to $\|y\| \leqslant \varepsilon$.*

Proof. Assume to the contrary that there is an arc y which satisfies

$$\|y\| \leqslant \varepsilon, \;\; y'(s) \in \widetilde{F}_s\big(y(s)\big) \text{ a.e.},$$

together with the boundary conditions, and for which $\ell(y(0), y(T)) < \ell(0,0)$. Define[2] an arc x on $[a,b]$ via $x(t) = y(\tau(t))$. It then follows that x satisfies

[2] This defines an arc because $y(\cdot)$ is absolutely continuous (by assumption) and $\tau(\cdot)$ is both absolutely continuous and strictly increasing.

$$\|x\| \leqslant \varepsilon\,,\;\; x'(t) \in F_t\big(x(t)\big) \text{ a.e.},$$

and that x satisfies the boundary conditions. But we have

$$\ell\big(x(a),x(b)\big) \,=\, \ell\big(y(0),y(T)\big) \,<\, \ell(0,0)\,,$$

which contradicts the optimality of $x_* \equiv 0$. □

It is clear that the transformed problem satisfies [H2] with $k \equiv 1$; thus, [TH1] holds. We may therefore apply the case of the theorem already proved; we deduce the existence of an arc $\widetilde{p}$ on the interval $[0,T]$ and η equal to 0 or 1 satisfying nontriviality, as well as the conditions (25.1 b) (25.1 c) (25.1 d) for $\widetilde{F}$.

We conclude by showing that the arc $p(t) = \widetilde{p}(\tau(t))$ on $[a,b]$ satisfies these same conditions for F; only (25.1 c) fails to be immediately apparent.

It follows directly from the definition of proximal normal that

$$(\alpha,\beta) \in N^P_{\widetilde{G}_s}(y,w) \iff \big(\alpha,\beta/k_t\big) \in N^P_{G_t}(y,k_t w)\,,$$

where $s = \tau(t)$. The equivalence also holds with $N^L_{\widetilde{G}_s}$ and $N^L_{G_t}$, by taking limits. Then, for almost every t:

$$\begin{aligned} p'(t) &= \widetilde{p}\,'\big(\tau(t)\big)\,\tau'(t) \,=\, \widetilde{p}\,'\big(\tau(t)\big)\,k_t \\ &\in \operatorname{co}\big\{\,\omega k_t : \big(\omega, \widetilde{p}(\tau(t))\big) \in N^L_{\widetilde{G}_{\tau(t)}}(0,0)\big\} \;\; \text{(by (25.1 c) for } \widetilde{p}) \\ &= \operatorname{co}\big\{\,\omega k_t : \big(\omega, p(t)\big) \in N^L_{\widetilde{G}_{\tau(t)}}(0,0)\big\} \\ &= \operatorname{co}\big\{\,\omega k_t : \big(\omega, p(t)/k_t\big) \in N^L_{G_t}(0,0)\big\} \;\; \text{(by the equivalence noted above)} \\ &= \operatorname{co}\big\{\,\omega : \big(\omega, p(t)\big) \in N^L_{G_t}(0,0)\big\}, \end{aligned}$$

since $N^L_{G_t}(0,0)$ is a cone. The proof of Theorem 25.1 is complete.

A corollary with running cost. For later use, we record a simple extension of Theorem 25.1 in which the cost in problem (DI) of p. 503 is modified as follows:

$$J(x) \,=\, \ell\big(x(a),x(b)\big) + \int_a^b \Lambda\big(t,x'(t)\big)\,dt\,.$$

We assume that Λ is measurable in t and globally Lipschitz in v (uniformly in t). The rest is unchanged.

25.4 Corollary. *There exist p and η as in the theorem, but satisfying the modified Euler inclusion*

$$p'(t) \in \operatorname{co}\big\{\,\omega : \big(\omega, p(t)\big) \in N^L_{G_t}\big(x_*(t),x_*'(t)\big) + \{0\}\times\eta\,\partial_L\Lambda\big(t,x_*'(t)\big)\big\} \text{ a.e.} \quad (20)$$

as well as the modified maximum condition, for almost every t:

$$\langle p(t), v\rangle - \eta\,\Lambda\big(t,v\big) \leqslant \langle p(t), x_*'(t)\rangle - \eta\,\Lambda\big(t, x_*'(t)\big) \;\; \forall\, v \in F_t\big(x_*(t)\big). \tag{21}$$

Proof. We absorb the running cost into an augmented problem, as follows. The state x gains a coordinate y, and becomes (x,y). The augmented multifunction, cost, and endpoint constraint set are given by

$$F_+(t,x,y) = \big\{ \big(v, \Lambda(t,v)\big) : v \in F_t(x) \big\}, \;\; \ell_+(x_0,y_0,x_1,y_1) = \ell(x_0,y_0) + y_1,$$
$$E_+ = \big\{ (x_0,y_0,x_1,y_1) : \big(x_0,x_1\big) \in E,\; y_0 = 0 \big\}.$$

It is easy to see that the augmented arc (x_*, y_*) is a local minimizer for the corresponding problem (DI), where

$$y_*(t) = \int_a^t \Lambda\big(s, x_*'(s)\big)\, ds,$$

and that the hypotheses of Theorem 25.1 are satisfied.

Upon applying the necessary conditions, there results η and an augmented costate arc (p,q). It follows that q is constant (from the Euler inclusion) and that $q = -\eta$ (from the transversality). Then the required nontriviality and transversality conditions are seen to hold, and the maximum condition of the augmented problem is precisely (21); only the Euler inclusion remains to be established.

Consider a point

$$\big(\omega, p, -\eta\big) \in N^L_{G_+}\big(x_*, v_*, \Lambda(v_*)\big),$$

where

$$G_+ = \big\{ (x,v,w) : v \in F(x),\; w - \Lambda(v) = 0 \big\}.$$

It is points such as these which figure in the augmented Euler inclusion (we have suppressed t, written v_* for $x_*'(t)$, and dropped the y variable, since F_+ does not depend on y). Note that G_+ may be expressed as the set of (x,v,w) satisfying $\varphi(x,v,w) \in \Phi$, where

$$\varphi(x,v,w) = \big(x, v, w - \Lambda(v)\big),\;\; \Phi = G \times \{0\}.$$

By Theorem 11.38, there exist $(\alpha,\beta) \in N^L_G(x_*,v_*)$ and $\lambda \in \mathbb{R}$ such that

$$\big(\omega, p, -\eta\big) \in \partial_L \big\{ \langle (\alpha,\beta), (x,v)\rangle + \lambda\,\big(w - \Lambda(v)\big) \big\} \big(x_*, v_*, \Lambda(v_*)\big).$$

It follows from this that $\lambda = -\eta$ and

$$(\omega, p) \in N^L_G(x_*,v_*) + \{0\} \times \eta\, \partial_L \Lambda(v_*).$$

The Euler inclusion (20) results. □

25.2 Proof of the extended maximum principle

In this section we derive the extended maximum principle (Theorem 22.26) from Theorem 25.1. The case $\Lambda \equiv 0$ is treated first.

A. We define Z to be the set of all pairs (η, p) where p is an arc on $[a,b]$, $\eta \geqslant 0$, $\|p\| + \eta = 1$, and where (η, p) satisfies the transversality condition (T) and the adjoint inclusion (A) of Theorem 22.26. Since Λ is zero, we have $H = H^1 = H^0$, and (A) has the form

$$-p'(t) \in \partial_C \langle p(t), f(t, \bullet, u_*(t)) \rangle \big(x_*(t)\big), \; t \in [a,b] \text{ a.e.} \tag{1}$$

Z is given the metric topology induced by the norm $\|(\eta, p)\| = |\eta| + \|p\|$.

Lemma 1. *Z is compact.*

Proof. For any $p \in \mathbb{R}^n$, for almost every t, the function $x \mapsto H^\eta(t, x, p, u_*(t))$ is Lipschitz near $x_*(t)$ of rank $k_*(t)|p|$, where $k_*(t) = k(t, u_*(t))$. It follows that any element ζ lying in the generalized gradient of this function at $x_*(t)$ is bounded: $|\zeta| \leqslant k_*(t)|p|$ (see Prop. 10.5). We deduce from this observation that any element $(\eta, p) \in Z$ satisfies, as a consequence of (1), the estimate

$$|p'(t)| \leqslant k_*(t)|p(t)| \text{ a.e.}$$

We use this to show that Z is sequentially compact. Let (η_i, p_i) be any sequence in Z. It follows from Gronwall's lemma and the estimate just derived (see Exer. 6.42) that (η_i, p_i) admits a subsequence (we do not relabel) such that $\eta_i \to \eta \in \mathbb{R}_+$, and such that p_i converges uniformly to an arc p, with p_i' converging weakly in $L^1(a,b)$ to p'. Clearly, (η, p) satisfies $\eta + \|p\| = 1$ as well as (T). To conclude, we need only check that (1) holds, for then (η, p) belongs to Z.

We prepare an appeal to the weak closure theorem 6.39. For each i, we have

$$-p_i'(t) \in \Gamma\big(t, p_i(t)\big), \; t \in [a,b] \text{ a.e.}, \tag{2}$$

where we define

$$\Gamma(t,p) = \partial_C H\big(t, \bullet, p, u_*(t)\big)\big(x_*(t)\big).$$

Then $\Gamma(t, \cdot)$ is convex-valued, and its graph is closed by a known property of the generalized gradient (see Prop. 10.10). We also have $\Gamma(t,p) \subset k_*(t)B$ whenever $|p| \leqslant 1$, as shown above. For any measurable function $q(t)$, the function

$$(t,x) \mapsto H\big(t, x, q(t), u_*(t)\big) = \langle q(t), f(t, x, u_*(t)) \rangle$$

is measurable in t (since f is LB measurable in t and u) and locally Lipschitz in x near $x_*(t)$. It follows that the multifunction $t \mapsto \Gamma(t, q(t))$ is measurable (see Exer. 13.24). In view of Prop. 6.29, this fact furnishes the final ingredient allowing us to invoke Theorem 6.39 and pass to the limit in (2). This yields (1). □

Now let $C = \{u_i(\cdot)\}_i$ be any *finite* collection of measurable functions u_i having the following properties:

(a) $u_* \in C$.

(b) For each i, we have $u_i(t) \in U(t)$ a.e.

(c) For some $\delta_C \in (0,1]$, we have, for each i, for almost every $t \in [a,b]$:

$$u_i(t) \neq u_*(t) \Longrightarrow \big| f\big(t, x_*(t), u_i(t)\big) - f\big(t, x_*(t), u_*(t)\big)\big| \geqslant \delta_C,$$
$$k\big(t, u_i(t)\big) \leqslant [1 + k(t, u_*(t))]/\delta_C.$$

We denote by $\mathscr{C}$ the set of all such collections C. Note that $\mathscr{C} \neq \emptyset$, since $\mathscr{C}$ contains the element $\{u_*(\cdot)\}$.

Lemma 2. *Let $C \in \mathscr{C}$. Then there exists an element $(\eta, p) \in Z$ such that, for every $u_i \in C$, we have*

$$\langle p(t), f\big(t, x_*(t), u_i(t)\big)\rangle \leqslant \langle p(t), f\big(t, x_*(t), u_*(t)\big)\rangle, \ t \in [a,b] \text{ a.e.}$$

Proof. The idea is to call upon Theorem 25.1 for a "reduced" optimal control problem defined in terms of the multifunction

$$F_t(x) = \big\{ f\big(t, x, u_i(t)\big) : u_i(\cdot) \in C \big\},$$

with ℓ and E unchanged. Hypothesis [H1] of Theorem 25.1 is satisfied in this context, as can be seen from the characterization

$$G_t = \big\{ (x,v) : \varphi(t,x,v) = 0 \big\}, \text{ where } \varphi(t,x,v) = \min_i \big| v - f\big(t, x, u_i(t)\big)\big|,$$

which shows that Prop. 6.25 applies (because φ is measurable in t and continuous in (x,v)). The Lipschitz condition in [H2] holds with $k_t = [1 + k(t, u_*(t))]/\delta_C$.

Let x be any trajectory for F which is admissible for the reduced problem, and satisfies $\|x - x_*\| \leqslant \varepsilon$. Then the multifunction

$$\Gamma(t) = \big\{ u \in \{u_i(t)\} : u_i(\cdot) \in C, \, x'(t) - f\big(t, x(t), u_i(t)\big) = 0 \big\},$$

being closed-valued and measurable, admits a measurable selection, from which it follows that x is an admissible state trajectory for the system (f, U). Thus we have $\ell(x(a), x(b)) \geqslant \ell(x_*(a), x_*(b))$, by the optimality of the process (x_*, u_*). Furthermore, since $u_* \in C$, the arc x_* is among the admissible trajectories for the reduced problem. We conclude that x_* provides a local minimum for the reduced problem, in the sense of Theorem 25.1.

We may therefore invoke Theorem 25.1 to deduce the existence of (η, p) satisfying $\eta + \|p\| = 1$ (this alternate form of nontriviality is explained in Prop. 22.5), the transversality condition (25.1 b), the maximum condition (25.1 d), and the Euler

inclusion (25.1 c) of that theorem. It is clear that the maximum condition yields the maximization property stated in the lemma. To conclude that $(\eta, p) \in Z$, we need only verify that the Euler inclusion for F implies the adjoint inclusion (1).

Let us consider any t and ω for which we have

$$\big(\omega, p(t)\big) \in N^L_{G_t}\big(x_*(t), x_*'(t)\big).$$

Observe that any point $(x,v) \in G_t$ sufficiently close to $(x_*(t), x_*'(t))$ is necessarily of the form

$$\big(x, f\big(t, x, u_*(t)\big)\big),$$

because of property (c) in the way C is defined. It follows that we have

$$\big(\omega, p(t)\big) \in N^L_S\big(x_*(t), x_*'(t)\big), \tag{3}$$

where S is defined by

$$S = \big\{ (x,v) : v - f\big(t, x, u_*(t)\big) = 0 \big\}.$$

We now prepare to apply Theorem 11.38. To this end, let $\varphi(x,v) = v - f(t, x, u_*(t))$. If, for some λ, we have

$$(0,0) \in \partial_L\langle \lambda, \varphi \rangle\big(x_*(t), x_*'(t)\big),$$

then we find $\lambda = 0$. This is the constraint qualification that allows us to invoke Theorem 11.38. We deduce that, as a consequence of (3), we have, for some λ:

$$p(t) = \lambda, \ \ \omega \in \partial_L\langle -\lambda, g \rangle\big(x_*(t)\big),$$

where g is defined by $g(x) = f(t, x, u_*(t))$. It follows that

$$\omega \in \partial_L\langle -p(t), g \rangle\big(x_*(t)\big) \subset \partial_C\langle -p(t), g \rangle\big(x_*(t)\big) = -\partial_C\langle p(t), g \rangle\big(x_*(t)\big).$$

Since, almost everywhere, $p'(t)$ is in the convex hull of such points ω (according to the Euler inclusion), and since $\partial_C g$ is convex-valued, we arrive at

$$-p'(t) \in \partial_C\langle p(t), g \rangle\big(x_*(t)\big),$$

which is precisely (1). Lemma 2 is proved. □

For any $C \in \mathscr{C}$, we denote by $M(C)$ the set of $(\eta, p) \in Z$ satisfying the conclusions of Lemma 2. Then $M(C)$ is a nonempty closed subset of Z. Observe that

$$M(C_1 \cup C_2) = M(C_1) \cap M(C_2).$$

Since $\mathscr{C}$ is evidently closed under finite unions, it follows from Lemma 2 that the family $\{M(C)\}_{C \in \mathscr{C}}$ has the finite intersection property. From the compactness of Z (Lemma 1), we infer the existence of an element (η, p) belonging to the intersection

of the entire family. We proceed to show that this element satisfies the maximum condition (M) of Theorem 22.26 almost everywhere.

We reason *ad absurdum*. If this is not the case, then there exists a subset S of $[a,b]$ of positive measure in which (M) fails. For any positive integer j, we set

$$U_j(t) = \big\{ u \in U(t) \text{ such that } \langle p(t), f\big(t, x_*(t), u\big) - f\big(t, x_*(t), u_*(t)\big)\rangle > 1/j,$$
$$\big| f\big(t, x_*(t), u\big) - f\big(t, x_*(t), u_*(t)\big)\big| > 1/j,\ k(t,u) < j\big(1 + k(t, u_*(t))\big)\big\}, \quad (4)$$

and we define S_j to be the set of $t \in [a,b]$ such that $U_j(t) \neq \emptyset$. Then $S = \cup_{j \geqslant 1} S_j$, so that, for some positive integer j, the set S_j has positive measure. It follows from the basic hypotheses, together with Prop. 6.36, that the graph of U_j is LB measurable. Thus, Aumann's selection theorem 23.3 yields the existence of a measurable function u_j having values in $U_j(t)$ for almost every $t \in S_j$; we define $u_j(t) = u_*(t)$ for $t \notin S_j$.

Then the collection C given by $\{u_*, u_j\}$ belongs to $\mathscr{C}$ (with $\delta_C = 1/j$). However, $(\eta, p) \notin M(C)$, since (M) fails on a set of positive measure. This is the required contradiction.

To obtain the (η, p) of the theorem, we need only normalize to obtain $\eta = 0$ or 1, essentially as explained in Prop. 22.5.

B. We now derive the constancy of the Hamiltonian asserted by Theorem 22.26 in the autonomous case, still with $\Lambda = 0$.

We introduce a new control component w with values in $[1-\delta, 1+\delta]$, a new state coordinate y, and the following augmented problem data:

$$f_+(x,y,u,w) = \big(w\, f(x,u), w\big),\ \ \ell_+(x_0, y_0, x_1, y_1) = \ell(x_0, x_1),$$
$$E_+ = \big\{ (x_0, y_0, x_1, y_1) : (x_0, x_1) \in E\,, y_0 = a,\ y_1 = b \big\},\ U_+ = U \times [1-\delta, 1+\delta].$$

We also define $w_* \equiv 1$ and $y_*(t) = t$. The positive number $\delta \in (0, 1/2)$ will be specified below. The reader will understand that the augmented process (x_*, y_*, u_*, w_*) is admissible for the augmented problem, and corresponds, in a certain sense, to the original process (x_*, u_*). We claim that it constitutes a local minimizer for the augmented process.

To see this, suppose to the contrary that some augmented process (x, y, u, w) with $\|x - x_*\| \leqslant \varepsilon/2$ is better; this translates as $\ell(x(a), x(b)) < \ell(x_*(a), x_*(b))$. As in the proof of Theorem 22.20 (see page 469), we induce a change of time scale via the bi-Lipschitz transformation

$$\tau(t) := a + \int_a^t w(\sigma)\, d\sigma\,.$$

Notice that τ increases from a to b as t does the same, since $y(b) - y(a) = b - a$. We proceed to define

$$x_+(\tau) = x\big(t(\tau)\big), \;\; u_+(\tau) = u\big(t(\tau)\big), \;\; \tau \in [a,b].$$

Then

$$\big|x_+(\tau) - x_*(\tau)\big| \leqslant \big|x\big(t(\tau)\big) - x_*\big(t(\tau)\big)\big| + \big|x_*\big(t(\tau)\big) - x_*(\tau)\big|.$$

If δ is chosen sufficiently small, the second term is guaranteed to be less than $\varepsilon/2$ for all $\tau \in [a,b]$; this results from the uniform continuity of the function x_* on $[a,b]$. So then we have $\|x_+ - x_*\| < \varepsilon$. Furthermore, we calculate

$$x'_+(\tau) = x'\big(t(\tau)\big)/w\big(t(\tau)\big) = f\big(x(t(\tau)), u(t(\tau))\big) = f\big(x_+(\tau), u_+(\tau)\big),$$

which shows that (x_+, u_+) is an admissible process for the original problem (the boundary conditions are clearly satisfied). We also have

$$\ell\big(x_+(a), x_+(b)\big) = \ell\big(x(a), x(b)\big) < \ell\big(x_*(a), x_*(b)\big),$$

which contradicts the optimality of (x_*, u_*). We conclude that the augmented process (x_*, y_*, u_*, w_*) is a local minimum for the augmented problem.

The hypotheses allow us to invoke the necessary conditions of Theorem 22.26 that were proved above: that is, we obtain (N), (T), (A), and (M) for the augmented problem. The augmented Hamiltonian is given by

$$H_+^\eta(x,y,p,q,u,w) = w\langle f(x,u), p\rangle + qw,$$

where we have denoted the extra costate coordinate by q. As regards the parts of the conclusion that pertain to p and u_*, it is easy to see that we recover (T), (A), and (M) for the original data. The (augmented) adjoint equation also provides $q'(t) = 0$ a.e., so that q is constant. The maximum condition with respect to w affirms that, almost everywhere, the function

$$w \mapsto w\big\{\langle p(t), f(x_*(t), u_*(t))\rangle + q\big\}$$

attains a maximum over $[1-\delta, 1+\delta]$ at $w = 1$. Thus, the coefficient of w equals 0 a.e. This yields precisely the constancy of the Hamiltonian (with $h = -q$).

There remains the nontriviality to verify. In augmented terms, we have the nontriviality condition $(\eta, p(t), q) \neq 0 \;\; \forall t$, but what we require is $(\eta, p(t)) \neq 0 \;\; \forall t$. Suppose to the contrary that $(\eta, p(\tau)) = 0$ for some τ; then $\eta = 0$ and q is a nonzero constant. Because the adjoint inclusion yields $|p'(t)| \leqslant k(t, u_*(t))|p(t)|$, Gronwall's lemma implies that p is identically zero. But then the equation

$$\langle p(t), f\big(x_*(t), u_*(t)\big)\rangle + q = 0 \text{ a.e.},$$

which was obtained above, cannot hold: a contradiction.

C. We now treat the case $\Lambda \neq 0$, by "absorbing" the running cost into the dynamics. We augment the state x by an additional coordinate y, and we define

$$f_+(t,x,y,u) = \big[f(t,x,u), \Lambda(t,x,u)\big], \;\; \ell_+(x_0,y_0,x_1,y_1) = \ell(x_0,x_1)+y_1,$$

$$E_+ = \big\{(x_0,y_0,x_1,y_1) : (x_0,x_1) \in E,\, y_0 = 0\big\},$$

$$x_+^*(t) = \Big[x_*(t), \int_a^t \Lambda\big(s,x_*(s),u_*(s)\big)\,ds\Big]$$

It is a notational exercise to check that, for the system (f_+,U), the (augmented) process (x_+^*, u_*) provides a local minimum for the cost

$$\ell_+\big(x(a),y(a),x(b),y(b)\big)$$

subject to $(x(a),y(a),x(b),y(b)) \in E_+$. Since this problem has zero running cost, and since (as is easily seen) the data satisfy the hypotheses, we may apply the case of Theorem 22.26 proved above. The Hamiltonian of the problem is

$$H_+^\eta(t,x,y,p,q,u) = H^{-q}(t,x,p,u),$$

where the additional costate coordinate has been labeled q. The (augmented) adjoint equation yields $q'(t) = 0$ a.e., since H_+^η does not depend on y; thus, q is constant. The augmented transversality condition, as it pertains to q, is $-q(b) = \eta$, so q is the constant $-\eta$. Then the costate p is seen to satisfy (T), (A), and (M).

Let us now verify nontriviality. If $(\eta, p(t)) = 0$ at some t, then $\eta = 0 = q$, and so $(\eta, p(t), q(t)) = 0$. This contradicts the (augmented) nontriviality condition. There remains the constancy of the Hamiltonian to prove, when the problem is autonomous. But in that case, the augmented Hamiltonian is autonomous too, and its constancy corresponds to the desired conclusion. □

Proof of Corollary 22.31.

Proof. The proof of Theorem 22.26 used the LB measurability of $U(\cdot)$ in just one step, in order to prove the existence of a measurable selection, on a suitable subset of S, of the multifunction defined by (4). (In this context, S is a set of positive measure on which (M) fails.) We now give an alternate argument that exploits the structural hypothesis 22.30 instead of LB measurability.

To ease the exposition, we give the proof only in the case of a partition of $[a,b]$ into two intervals, $[a,c]$ and $[c,a]$. (The argument extends easily to any finite or countable partition.) Then either $S \cap (a,c)$ or $S \cap (c,b)$ has positive measure; let us consider the first case.

Let $\{v_i\}$ be a countable dense subset of U_1, where $U(t) = U_1$ for $t \in [a,c)$. For positive integers i and j, define

$$S_{i,j} = \big\{t \in (a,c) \text{ such that } \langle p(t), f\big(t,x_*(t),v_i\big) - f\big(t,x_*(t),u_*(t)\big)\rangle > 1/j,$$
$$\big|f\big(t,x_*(t),v_i\big) - f\big(t,x_*(t),u_*(t)\big)\big| > 1/j, \;\; k(t,v_i) < j\big(1+k(t,u_*(t))\big)\big\}.$$

One may verify that this defines a measurable set. Then, because of the continuity of f in u, we have $S \cap (a,c) = \cup_{i,j \geqslant 1} S_{i,j}$, so that some $S_{i,j}$ has positive measure. We then define a control u_j that equals v_i on $S_{i,j}$, and u_* elsewhere. From this point, the proof concludes as before. □

Remark. The proof of Cor. 22.31 does not call upon the Aumann selection theorem 23.3. The use of the latter in proving Theorem 22.26 may also be avoided another way, by positing that $U(\cdot)$ is measurable and closed-valued (which is a stronger hypothesis than LB measurability).

25.3 Stratified necessary conditions

In considering a control system given in the form of a differential inclusion

$$x'(t) \in F_t\big(x(t)\big),$$

certain situations arise in which the Lipschitz assumption [H2] of Theorem 25.1 is inappropriate. An example is provided by an optimal control problem based on a standard control system (f,U), but which features additional *mixed constraints* involving both the state and the control, such as a condition of the type

$$g\big(t,x(t),u(t)\big) \leqslant 0.$$

The question of finding an equivalent differential inclusion is not affected by this extra constraint: we simply define

$$F_t(x) = \big\{ f(t,x,u) : u \in U,\ g(t,x,u) \leqslant 0 \big\}.$$

In contrast to when the new constraint was absent, however, it is now unrealistic to expect F to be Lipschitz in x under any reasonable set of assumptions.

Instead, a certain localized pseudo-Lipschitz property can be expected to hold, if the mixed constraint is non degenerate. For dealing with such a situation, we introduce below a localization device called a *radius*, and a constraint qualification called the *bounded slope condition*.[3] The resulting theorem on differential inclusions, the final stage in the edifice of necessary conditions that we have been building, will be the key to deriving the multiplier rule in optimal control, as well as in the calculus of variations.

The problem we discuss in this section is identical in appearance to the one treated earlier by Theorem 25.1 (it is the hypotheses that are different):

[3] The reader has seen the phrase "bounded slope condition" used before once or twice, but we believe the context will preclude any confusion.

$$\begin{cases} \text{Minimize} & J(x) = \ell\big(x(a), x(b)\big) \\ \text{subject to} & x'(t) \in F_t\big(x(t)\big),\ t \in [a,b] \text{ a.e.} \\ & \big(x(a), x(b)\big) \in E. \end{cases} \qquad \textbf{(DI)}$$

We are also given an arc x_* that is admissible for (DI).

Now let R be a multifunction from $[a,b]$ to $\mathbb{R}^n$ such that $x_*'(t) \in R(t)$ a.e. We say that x_* is a *local minimizer of radius* R for the problem provided that, for some $\varepsilon > 0$, for all admissible arcs x satisfying $\|x - x_*\| \leqslant \varepsilon$ as well as

$$x'(t) \in F_t\big(x(t)\big) \cap R_t \text{ a.e.},$$

we have $J(x_*) \leqslant J(x)$. The multifunction R will be called a *radius* (around x_*). This terminology is inspired by the important special case in which R_t is a ball centered at $x_*'(t)$ for each t, but this is not assumed here. In a sense, the introduction of the radius allows us to consider a type of *weak* (rather than strong) local minimum.

As before (§25.1), the notation G_t below refers to the graph of the multifunction $F_t(\cdot)$. The following hypotheses will be involved.

[H3] The function ℓ is locally Lipschitz; the set E is closed; the multifunction $t \mapsto G_t$ is measurable; the following set is closed for almost every t:

$$\big\{ (x,v) \in G_t : |x - x_*(t)| \leqslant \varepsilon \big\}.$$

[H4] (bounded slope condition) There exists a summable function k such that, for almost every t, the following implication holds:

$$|x - x_*(t)| \leqslant \varepsilon,\ v \in F_t(x) \cap R_t,\ (\alpha, \beta) \in N^P_{G_t}(x,v) \implies |\alpha| \leqslant k_t |\beta|.$$

[H5] The multifunction $t \mapsto R_t$ is measurable; for some $\delta > 0$, for almost every t, the set R_t is open and convex and satisfies

$$R_t \supset B\big(x_*'(t), \delta\, k_t\big).$$

The reader will observe that [H3] essentially restates [H1] of Theorem 25.1; it is included here for completeness.

The theorem below is referred to as *stratified* because the hypotheses (including the optimality) are assumed to hold relative to a different set R_t for each t, and the conclusions are asserted to precisely that extent (that is, on R_t).

25.5 Theorem. *Let x_* be a local minimizer of radius R for the problem* (DI), *under hypotheses* [H3][H4][H5] *above. Then there exist an arc p and a scalar η equal to* 0 *or* 1 *satisfying the* **nontriviality condition**

$$\big(\eta, p(t)\big) \neq 0 \ \ \forall t \in [a,b], \qquad (25.5\text{ a})$$

the **transversality condition**

$$\big(p(a), -p(b)\big) \in \eta\,\partial_L \ell\big(x_*(a), x_*(b)\big) + N^L_E\big(x_*(a), x_*(b)\big)\,, \tag{25.5 b}$$

the **Euler inclusion** *for almost every* t:

$$p'(t) \in \operatorname{co}\big\{\,\omega : \big(\omega, p(t)\big) \in N^L_{G_t}\big(x_*(t), x_*'(t)\big)\big\} \text{ a.e.} \tag{25.5 c}$$

as well as the **maximum condition** *of radius* R *for almost every* t:

$$\langle\, p(t), v\,\rangle \leqslant \langle\, p(t), x_*'(t)\,\rangle \;\; \forall\, v \in F_t\big(x_*(t)\big) \cap R_t\,. \tag{25.5 d}$$

Remarks.

(a) It can be shown that when R_t is identically $\mathbb{R}^n$, the theorem above is equivalent to Theorem 25.1 (which is actually used in proving Theorem 25.5). One cannot obtain Theorem 25.5 by simply replacing F_t by $F_t \cap R_t$ and then applying Theorem 25.1, however; the intersection will lack the required Lipschitz property.

(b) The bounded slope condition [H4] may be replaced in the theorem statement by the following explicit pseudo-Lipschitz hypothesis:

$$x, y \in B(x_*(t), \varepsilon) \implies F_t(x) \cap R_t \subset F_t(y) + B(0, k_t|x-y|)\,.$$

We shall establish below in the course of events that the bounded slope condition implies (essentially) this property, which is what is used subsequently in the proof of the theorem.

(c) The theorem fails in the absence of the δ of hypothesis [H5]: surprisingly perhaps, the radius (and hence the extent of the optimality) must be big enough with respect to the function k that calibrates the pseudo-Lipschitz behavior of F.[4]

A reduction. We may (and do) assume without loss of generality that the solution x_* is identically zero. Indeed, one can redefine the data of the problem by translation to attain that situation: $F_t(x)$ is replaced by $F_t(x_*(t)+x) - x_*'(t)$, R_t by $R_t - x_*'(t)$, etc. It is easy to see that both the hypotheses and the conclusions of the theorem are robust relative to this normalization.

Temporary hypotheses. Extra hypotheses strengthening [H4] and [H5] will now be made; their removal will be the final step in the proof.

[TH1] The function k is a positive constant.

[TH2] For some $M > \delta k > 0$, for almost every t, we have

$$B(0, \delta k) \subset R_t \subset B(0, M)\,.$$

[4] An appropriate counter-example is given in [16].

We see from [H5] and [TH2] that R_t is a *bounded* convex body containing the ball $B(0,\delta k)$; we denote by g_t its gauge function (see Theorem 2.36):

$$g_t(f) = \inf\{\lambda > 0 : f/\lambda \in R_t\}.$$

Then $R_t = \{f : g_t(f) < 1\}$, and $g_t(f) \leqslant |f|/(\delta k)$. It follows from subadditivity that g_t has global Lipschitz constant $1/(\delta k)$, and it is not difficult to see that the mapping $t \mapsto g_t(f)$ is measurable[5] for each f.

A Lipschitz lifting. We fix $r > 0$ sufficiently small so that

$$\delta r < \min(\varepsilon, \delta), \;\; r < (1-r)^2, \;\; r(2-r) < 1/4, \;\; r < 1/4 \tag{1}$$

and we fix any t for which [H3] [H4] [H5] [TH2] hold. The following type of property has been called *pseudo-Lipschitz*.

25.6 Proposition. *Let x_1, x_2 belong to $B(0,\delta r/3)$, and let $f_1 \in F_t(x_1)$ satisfy $g_t(f_1) \leqslant 1-r$. Then there exists $f_2 \in F_t(x_2)$ satisfying $|f_2 - f_1| \leqslant k|x_2 - x_1|$.*

Proof. Let $C = \{(x,v) \in G_t : |x| \leqslant \varepsilon,\; g_t(v) \leqslant 1 - r/5\}$. Then [H3] implies that C is compact. We shall now apply the mean value inequality (Theorem 11.2) with the data

$$f(y,w) = I_C(y,w) + |w - f_1|, \;\; x = (x_1, f_1), \; Y = \{x_2\} \times B\big(f_1, k|x_2 - x_1|\big),$$

and for any real number $\bar r$ satisfying

$$\begin{aligned} \bar r &< \min\big\{|v - f_1| : (x_2, v) \in C\big\} \\ &\leqslant \min\big\{|v - f_1| : (x_2, v) \in C,\, |v - f_1| \leqslant k|x_2 - x_1|\big\} \\ &= \min_Y f - f(x). \end{aligned}$$

We do *not* assert this last quantity to be finite at this point, since we cannot exclude yet the possibility that f equals ∞ on Y. We deduce from Theorem 11.2, for any $\rho > 0$, the existence of $(z,u) \in \operatorname{dom} f$ and $(\zeta,\psi) \in \partial_P f(z,u)$ such that

$$(z,u) \in [x,Y] + \rho B = \operatorname{co}\big[\{(x_1,f_1)\} \cup \{x_2\} \times B\big(f_1, k|x_2 - x_1|\big)\big] + \rho B \tag{2}$$

$$\bar r \leqslant \langle (\zeta,\psi), (x_2 - x_1, w - f_1)\rangle \;\; \forall w \in B\big(f_1, k|x_2 - x_1|\big). \tag{3}$$

Note that $(z,u) \in C$ necessarily; by the proximal sum rule (Theorem 11.16), we may write

$$(\zeta,\psi) = (\alpha,\beta) + (0,\theta) \tag{4}$$

for some $(\alpha,\beta) \in N_C^L(z,u)$ and vector $\theta \in B(0,1)$.

[5] One may show this, for example, by proving that the multifunction $\Gamma(t) = \operatorname{cl} R_t$ is measurable, and then using a representation for Γ of the type furnished by Theorem 6.22.

We see from (2) that the distance from z to the segment $[x_1, x_2]$ is no greater than ρ, which, for ρ chosen sufficiently small, implies $|z| < \varepsilon$, in light of the first inequality in (1) (and since $x_1, x_2 \in B(0, \delta r/3)$). Also, we have $|u - f_1| \leqslant k|x_2 - x_1| + \rho$ (by (2) again) and $|x_2 - x_1| \leqslant 2\delta r/3 < \varepsilon$, so that

$$|u - f_1| \leqslant 2k\delta r/3 + \rho.$$

We derive, for ρ sufficiently small,

$$g_t(u) \leqslant g_t(f_1) + (\delta k)^{-1}|u - f_1| \leqslant 1 - r + 2r/3 + \rho/(\delta k) \leqslant 1 - r/4.$$

It follows therefore that C and G_t coincide near (z, u), whence $(\alpha, \beta) \in N^L_{G_t}(z, u)$ and $|\alpha| \leqslant k|\beta|$, by [H4]. (We use here the fact that the bounded slope condition continues to hold when $N^P_{G_t}(x, v)$ is replaced by $N^L_{G_t}(x, v)$; this results from an evident limiting argument.)

Choosing in (3) the available w that minimizes the right-hand side, and substituting (4), we discover

$$\begin{aligned} \bar{r} &\leqslant \big\{ |\alpha| - k|\beta + \theta| \big\} |x_2 - x_1| \\ &\leqslant \big\{ |\alpha| - k|\beta| \big\} |x_2 - x_1| + k|\theta||x_2 - x_1| \\ &\leqslant k|x_2 - x_1|. \end{aligned}$$

This implies that $\bar{r}$ is bounded above; more to the point, since $\bar{r}$ was chosen to be *any* number less than $\min\{|v - f_1| : (x_2, v) \in C\}$, we deduce

$$\min \big\{ |v - f_1| : (x_2, v) \in C \big\} \leqslant k|x_2 - x_1|.$$

This implies the existence of f_2 with $(x_2, f_2) \in C$ such that $|f_2 - f_1| \leqslant k|x_2 - x_1|$. As shown above (in connection with u), this inequality implies $g_t(f_2) \leqslant 1 - r/4$. Thus C and G_t coincide in a neighborhood of (x_2, f_2), and we have $f_2 \in F_t(x_2)$ as desired. □

The following *lifting* of F will play a central role:

$$F^r_t(x) = \big\{ (\lambda f, \lambda) : \lambda \in [0,1],\ f \in F_t(x),\ g_t(f) \leqslant (1 - \lambda r)(1 - r) \big\}. \quad (5)$$

Note that the values of f involved here all lie in R_t, and hence are bounded in norm by M, in view of [TH2]. We set

$$k^r := k + \frac{M+1}{\delta r(1-r)}.$$

25.7 Proposition. *F^r_t is Lipschitz in the following sense:*

$$y_1, y_2 \in B(0, \delta r/3) \implies F^r_t(y_1) \subset F^r_t(y_2) + B\big(0, k^r|y_2 - y_1|\big).$$

Proof. Let $(\lambda_1 f_1, \lambda_1) \in F_t^r(y_1)$ be given, and set $\rho = |y_2 - y_1|$. We must exhibit $(\lambda f, \lambda) \in F_t^r(y_2)$ satisfying

$$\big|(\lambda f, \lambda) - (\lambda_1 f_1, \lambda_1)\big| \leqslant k^r \rho. \tag{6}$$

Case 1. $\lambda_1 \leqslant \rho/[\,\delta r(1-r)]$.

We apply Prop. 25.6 with $x_1 = 0$, $f_1 = 0$ to see that some $f \in F_t(y_2)$ satisfies

$$|f| \leqslant k|y_2| < k\delta r,$$

whence $f/r \in R_t$ (by [TH2]) and $g_t(f) \leqslant r \leqslant (1-r)$ (by the second inequality in (1)), so that the point $(0f, 0)$ lies in $F_t^r(y_2)$ by definition. We proceed to verify (6) for this choice:

$$\big|(\lambda_1 f_1, \lambda_1)\big| \leqslant \lambda_1\big(1 + |f_1|\big) \leqslant \rho(1+M)/[\,\delta r(1-r)] \leqslant k^r \rho.$$

Case 2. $\lambda_1 > \rho/[\,\delta r(1-r)]$.

Now we set $\lambda = \lambda_1 - \rho/[\,\delta r(1-r)]$, and we invoke Prop. 25.6 to deduce the existence of $f \in F_t(y_2)$ satisfying $|f - f_1| \leqslant k|y_2 - y_1|$. Then

$$\begin{aligned} g_t(f) &\leqslant g_t(f_1) + (\delta k)^{-1}|f - f_1| \\ &\leqslant (1-\lambda_1 r)(1-r) + (\delta k)^{-1} k\rho \\ &= (1-\lambda r)(1-r). \end{aligned}$$

Thus $(\lambda f, \lambda) \in F_t^r(y_2)$ by definition. We conclude by checking (6):

$$\begin{aligned} \big|(\lambda f, \lambda) - (\lambda_1 f_1, \lambda_1)\big| &\leqslant \lambda_1 - \lambda + |\lambda f - \lambda_1 f| + |\lambda_1 f - \lambda_1 f_1| \\ &\leqslant (\lambda_1 - \lambda)(1 + |f|) + |\lambda_1||f - f_1| \\ &\leqslant (\lambda_1 - \lambda)(1 + |f|) + k\rho \\ &\leqslant \big\{\rho/[\,\delta r(1-r)]\big\}\{1+M\} + k\rho = k^r \rho. \end{aligned}$$ ☐

An auxiliary problem. For a positive sequence ε_i decreasing to 0, we define

$$\begin{aligned} \ell_i\big(x(a), x(b)\big) &= \big[\ell\big(x(a), x(b)\big) - \ell(0,0) + \varepsilon_i^2\,\big]_+ \\ \Lambda_t(v, \lambda) &= [g_t(v) - \lambda(1-\lambda r)(1-r) + \lambda r]_+ \\ \widetilde{E} &= \big\{(x,y,z,w) \in \mathbb{R}^n \times \mathbb{R} \times \mathbb{R}^n \times \mathbb{R} : (x,z) \in E,\ y = a\big\}. \end{aligned}$$

We let A be the set of arcs (x,y) satisfying $(x', y') \in F_t^r(x,y)$ a.e., as well as

$$\big(x(a), y(a), x(b), y(b)\big) \in \widetilde{E},\ |x(t)| \leqslant \delta r/3\ \ \forall t. \tag{7}$$

Here we have written $F_t^{\,r}(x,y)$ for $F_t^{\,r}(x)$, though there is no y-dependence; note that $\delta r < \varepsilon$ in light of (1). We now consider the problem of minimizing over A the cost

$$J_i(x,y) = \ell_i\big(x(a),x(b)\big) + b - y(b) + \int_a^b \Lambda_t\big(x'(t),y'(t)\big)\,dt\,.$$

Note that for any (x,y) that is admissible for this problem, we have $y(b) \leqslant b$. Thus, the infimum in the above problem is nonnegative (since $\ell_i \geqslant 0,\ \Lambda_t \geqslant 0$).

If any such (x,y) has $y(b)=b$, then $y'=1$ a.e., and it follows that x is a trajectory for F in the class relative to which x_* is optimal. Thus $\ell_i\big(x(a),x(b)\big) > 0$ (or else the optimality of x_* would be contradicted). We summarize as follows: for any (x,y) in A, letting $(x',y') = (\lambda f,\lambda)$, we have either $\lambda < 1$ on a set of positive measure, or else $\ell_i(x(a),x(b)) > 0$.

Note that the arc $(0,t)$ lies in A and yields $J_i = \varepsilon_i^2$ (since $\Lambda_t(0,1)=0$ by the second inequality in (1)). Thus $(0,t)$ is ε_i^2-optimal for the problem (whose infimum may not be attained). We define a complete metric d on the set A as follows:

$$d\big((x_1,y_1),(x_2,y_2)\big) = \\ \big|x_1(a)-x_2(a)\big| + \int_a^b \big|x_1'(t)-x_2'(t)\big|\,dt + \int_a^b \big|y_1'(t)-y_2'(t)\big|\,dt.$$

By Theorem 5.19 there exists $(x_i,y_i) \in A$ satisfying

$$\big|x_i(a)\big| + \int_a^b \big|x_i'(t)\big|\,dt + \int_a^b \big|y_i'(t)-1\big|\,dt \leqslant \varepsilon_i \tag{8}$$

and such that the minimum over A of the functional

$$J_i(x,y) + \varepsilon_i\big|x(a)-x_i(a)\big| + \varepsilon_i\int_a^b \big|x'(t)-x_i'(t)\big|\,dt + \varepsilon_i\int_a^b \big|y'(t)-y_i'(t)\big|\,dt$$

is attained at $(x_i,\,y_i)$. We denote y_i' by λ_i. It is a straightforward exercise in measurable selection theory (using Cor. 6.23) to show that we may write $x_i'(t)$ in the form $\lambda_i(t)f_i(t)$ for some measurable function f_i satisfying

$$f_i(t) \in F_t(x_i(t)) \text{ a.e.}, \quad g_t\big(f_i(t)\big) \leqslant \big(1-r\lambda_i(t)\big)(1-r) \text{ a.e.}$$

It follows from the above that we have

$$\|x_i\| \to 0 \text{ and, in } L^1(a,b)\text{: } x_i' \to 0,\ \lambda_i \to 1,\ f_i \to 0.$$

By taking subsequences, the last three convergences may be assumed to hold in the pointwise (almost everywhere) sense as well. Let us define

$$\begin{aligned} G_t^r &= \big\{(x,\lambda f,\lambda) : (\lambda f,\lambda) \in F_t^{\,r}(x)\big\} \\ &= \big\{(x,\lambda f,\lambda) : \lambda \in [0,1],\ f \in F_t(x),\ g_t(f) \leqslant (1-\lambda r)(1-r)\big\}, \end{aligned}$$

a set whose intersection with $B(0,\varepsilon)\times\mathbb{R}^n$ is closed a.e., as a result of [H3]. It also follows (with the help of Prop. 6.25) that the multifunction $t\mapsto G_t^r$ is measurable.

25.8 Proposition. *For all i sufficiently large, there exist an arc p_i, a nonnegative constant q_i and a measurable function (d_i,e_i) such that $\|p_i\|+q_i = 1$ and the following hold:*

$$\big(p_i(a),-p_i(b)\big) \in \partial_L q_i \ell_i\big(x_i(a),x_i(b)\big)+N_E^L\big(x_i(a),x_i(b)\big)+\varepsilon_i q_i B\times\{0\}, \quad (9)$$

$$p_i' \in \operatorname{co}\big\{\,\omega : (\omega, p_i-d_i, q_i-e_i) \in N^L_{G_t^r}\big(x_i,x_i',y_i'\big)\big\} \text{ a.e.}, \quad (10)$$

where

$$\big(d_i(t),e_i(t)\big)\in\partial_L q_i\Lambda_t\big(x_i'(t),y_i'(t)\big)+\varepsilon_i q_i B\times\varepsilon_i q_i B \text{ a.e.}$$

Further, the following maximum condition holds at almost every t:

$$(\lambda f,\lambda)\in F_t^r(x_i) \implies$$
$$\langle p_i,\lambda f\rangle+q_i\lambda-q_i\Lambda_t(\lambda f,\lambda)-\varepsilon_i q_i\big|\lambda f-\lambda_i f_i\big|-\varepsilon_i q_i|\lambda-\lambda_i|$$
$$\leqslant \langle p_i,\lambda_i f_i\rangle+q_i\lambda_i-q_i\Lambda_t(\lambda_i f_i,\lambda_i). \quad (11)$$

Proof. For i sufficiently large, (x_i,y_i) solves a problem (locally) to which Cor. 25.4 applies: hypothesis [H2] follows from Prop. 25.7. We deduce the existence of an arc (p_i,q_i) satisfying the necessary conditions given there, for some η_i in $\{0,1\}$. It follows that $q_i\equiv\eta_i$, and we get precisely the conditions (9) (10) (11). We then redefine the multipliers p_i and q_i by normalizing (dividing by $\|p\|+q_i\neq 0$); this leaves the preceding conditions unchanged, but provides the nontriviality in the alternate form $\|p\|+q_i = 1$ (which turns out to be more convenient for later convergence steps). □

Let Ω_i denote the subset of points t in $[a,b]$ for which we have

$$g_t\big(x_i'(t)\big)-\lambda_i(t)\big(1-\lambda_i(t)\,r\big)(1-r)+r < 0.$$

Note that the left side converges almost everywhere to $r-(1-r)^2<0$ (by (1)); it follows that meas $\Omega_i\to b-a$. The following refines the information we have about the multipliers identified above.

25.9 Proposition. *For all i sufficiently large, the arc p_i satisfies*

$$\big|p_i'(t)\big| \leqslant k^r\big\{\,|p_i|+(k\delta)^{-1}+3\,\big\},\ t\in[a,b] \text{ a.e.} \quad (12)$$

In addition, there exists $\gamma_i\in[0,1]$ such that p_i satisfies

$$\big(p_i(a),-p_i(b)\big) \in \partial_L\gamma_i q_i\ell\big(x_i(a),x_i(b)\big)+N_E^L\big(x_i(a),x_i(b)\big)+\varepsilon_i q_i B\times\{0\}, \ (13)$$

where $\gamma_i = 1$ if $\ell_i(x_i(a),x_i(b))>0$. We also have, for almost every $t\in\Omega_i$,

$$p_i' \in \operatorname{co}\big\{\,\omega : (\omega, p_i, q_i) \in N^L_{G_t^r}\big(x_i, x_i', y_i'\big) + \{0\}\times\varepsilon_i q_i B\times\varepsilon_i q_i B\,\big\}, \tag{14}$$

and, for almost every $t \in [a,b]$,

$$f \in F_t\big(x_i(t)\big),\; g_t(f) \leqslant (1-r)^2 - r \implies \\ \langle\, p_i, f\,\rangle \leqslant \langle\, p_i, \lambda_i f_i\,\rangle + \varepsilon_i\,|\,f - \lambda_i f_i\,| + \varepsilon_i\,|1-\lambda_i\,|. \tag{15}$$

Proof. It is a simple exercise to show that $|\,D_v \Lambda_t\,|$ and $|\,D_\lambda \Lambda_t\,|$ are bounded by $(k\delta)^{-1}$ and 1 respectively whenever these derivatives exist (the second estimate uses (1)). It follows that $|\,d_i\,| \leqslant (k\delta)^{-1} + \varepsilon_i$ and $|\,e_i\,| \leqslant 1 + \varepsilon_i$. The Lipschitz property of F_t^r provided by Prop. 25.7 implies by Prop. 25.2 that any ω as described in (10) satisfies

$$\begin{aligned}|\,\omega\,| &\leqslant k^r\big\{\,|\,p_i - d_i\,| + |\,q_i - e_i\,|\,\big\} \\ &\leqslant k^r\big\{\,|\,p_i\,| + [\,(k\delta)^{-1} + 1\,] + 2\,\varepsilon_i + 1\,\big\} \leqslant k^r\big\{\,|\,p_i\,| + (k\delta)^{-1} + 3\,\big\}\end{aligned}$$

for i large enough to imply $\varepsilon_i < 1$. We obtain (12).

The proximal chain rule (Theorem 11.41) provides the estimate

$$\partial_L \ell_i\big(x_i(a), x_i(b)\big) \subset \gamma_i\,\partial_L \ell\big(x_i(a), x_i(b)\big), \quad \gamma_i \in [0,1],$$

so (13) follows from (9). If $\ell_i(x_i(a), x_i(b)) > 0$, then ℓ_i and ℓ differ by a constant locally around the point $(x_i(a), x_i(b))$, whence

$$\partial_L \ell_i\big(x_i(a), x_i(b)\big) = \partial_L \ell\big(x_i(a), x_i(b)\big),$$

so that (13) holds with $\gamma_i = 1$.

When $t \in \Omega_i$, the relation (14) is a consequence of (10), since then Λ_t is identically zero near $(x_i'(t), y_i'(t))$ (as follows from its definition), so that

$$\partial_L\, q_i \Lambda_t\big(x_i'(t), y_i'(t)\big) = \{(0,0)\}.$$

Finally, let us derive (15). When f is as stated, then the point $(f,1)$ belongs to $F_t^r(x_i(t))$, and we have $\Lambda_t(f,1) = 0$. Invoking (11), we obtain

$$\langle\, p_i, f\,\rangle \leqslant \langle\, p_i, \lambda_i f_i\,\rangle + q_i(\lambda_i - 1) + \varepsilon_i q_i\,|\,f - \lambda_i f_i\,| + \varepsilon_i q_i\,|1-\lambda_i\,| - q_i \Lambda_t(\lambda_i f_i, \lambda_i),$$

which yields (15), since $q_i \in [0,1]$ and $\Lambda_t \geqslant 0$. □

Convergence. In light of the above, we may take subsequences as necessary (without relabeling) in order to arrange

$$q_i \to q,\;\; \gamma_i \to \gamma,\;\; p_i \to p \text{ (uniformly)},\;\; p_i' \to p' \text{ (weakly in } L^1(a,b)).$$

(This now-familiar argument uses (12) in conjunction with Gronwall's lemma, see Exer. 6.42.) Let S_i be the subset of $[a,b]$ for which $\lambda_i(t) < 1$. As remarked previously, when S_i has measure 0, it follows that $\ell_i(x_i(a), x_i(b))$ is strictly positive. By taking further subsequences, we can therefore arrange for one of the three cases treated below to arise:

Case 1 : $q_i \leqslant 1/2 \ \forall i$

Case 2 : $\ell_i(x_i(a), x_i(b)) > 0 \ \forall i$

Case 3 : $q_i > 1/2$, meas $S_i > 0 \ \forall i$

We now pass to the limit in (13) (14) (15) in order to obtain an arc p and nonnegative numbers q, γ such that $\|p\| + q = 1$ and

$$\big(p(a), -p(b)\big) \in \partial_L \gamma q\, \ell(0,0) + N_E^L(0,0) \tag{16}$$

$$p'(t) \in \operatorname{co}\big\{\omega : \big(\omega, p(t), q\big) \in N^L_{G_t^r}(0,0,1)\big\},\ t \in [a,b] \text{ a.e.} \tag{17}$$

$$f \in F_t(0),\ g_t(f) \leqslant (1-r)^2 - 2r \implies \langle p, f \rangle \leqslant 0. \tag{18}$$

Relation (16) results from the closed graph property of the limiting constructs. Since F_t^r is Lipschitz (by Prop. 25.7), Prop. 25.2 implies that G_t^r satisfies the bounded slope condition; we have seen that $t \mapsto G_t^r$ is closed-valued and measurable. Then (17) follows from (14), with the help of Prop. 25.3.

To see how (18) follows, consider any f as described there, where t is such that all hypotheses and pointwise convergences hold, as well as (15) for every i. By Prop. 25.6, for all large i, there exists $f_i' \in F_t(x_i(t))$ such that $|f_i' - f| \leqslant k|x_i(t)| < r\delta k$. Then

$$g_t(f_i') \leqslant g_t(f) + |f_i' - f|/(\delta k) < (1-r)^2 - 2r + r = (1-r)^2 - r.$$

We may therefore apply (15) to discover

$$\langle p_i, f_i' \rangle \leqslant \langle p_i, \lambda_i f_i \rangle + \varepsilon_i |f_i' - \lambda_i f_i| + \varepsilon_i |1 - \lambda_i|.$$

Passing to the limit (recall that $f_i \to 0$ a.e.) gives $\langle p, f \rangle \leqslant 0$, whence (18).

The conclusions (16) (17) (18) resemble closely the ones we seek to obtain in the theorem (although the maximum condition is not yet asserted to the required extent). But we need to address the issue of nontriviality by ruling out the possibility that γq and p are both zero. This is done differently in accord with which of the three cases is considered.

In Case 1, the inequality $q_i \leqslant 1/2$ implies $\|p_i\| \geqslant 1/2$, since $\|p_i\| + q_i = 1$, so the arc p above is nonzero.

In Case 2, we have (by Prop. 25.9) $\gamma_i = 1$. Thus $\gamma q + \|p\| = q + \|p\| = 1$.

Case 3, in which $q_i > 1/2$ and meas $S_i > 0 \;\; \forall\, i$, is treated in two further subcases. We shall prove that for all i sufficiently large,

$$\|\, p_i \,\| \geqslant \min\big\{1/(5M),\, r/(12M)\big\}. \tag{19}$$

For each i, there is a point t in S_i for which (11) holds (since S_i has positive measure). We fix such a t (suppressing it in the notation). Then one of the two alternatives below is valid:

Case 3(a) : $g_t(f_i) < (1-\lambda_i r)(1-r)$

In this case, for all $\lambda > \lambda_i$ sufficiently near λ_i, the point $(\lambda f_i, \lambda)$ lies in $F_t^{\,r}(x_i)$, so that we may invoke (11) for such values of λ. It follows from the presence of a maximum that the derivative Δ from the right with respect to the variable λ (evaluated at $\lambda = \lambda_i$) of the left side in the inequality appearing in (11), must be nonpositive. To calculate this derivative as regards the term involving Λ_t, observe that

$$\Lambda_t(\lambda f_i, \lambda) = \lambda\big\{\, g_t(f_i) - (1-\lambda r)(1-r) + r \,\big\}_+,$$

since g_t is positively homogeneous. If the quantity $Q(\lambda)$ in the braces above is strictly negative at $\lambda = \lambda_i$, the derivative there is zero; otherwise (by the product rule), the derivative at $\lambda = \lambda_i$ equals $Q(\lambda_i) + \lambda_i r(1-r) \leqslant r + \lambda_i r(1-r)$. In either case we obtain, from the inequality $\Delta \leqslant 0$:

$$\begin{aligned}\langle\, p_i, f_i \,\rangle &\leqslant -q_i + r + \lambda_i r(1-r) + \varepsilon_i(1+M)\\ &\leqslant -q_i + r(2-r) + \varepsilon_i(1+M)\\ &< -1/2 + r(2-r) + \varepsilon_i(1+M) < -1/4 + \varepsilon_i(1+M),\end{aligned}$$

since $r(2-r) < 1/4$ by (1). Because we have $|\, f_i \,| \leqslant M$, we conclude that, provided ε_i satisfies $\varepsilon_i(1+M) \leqslant 1/20$, then $|\, p_i \,| \geqslant 1/(5M)$, which confirms (19).

Case 3(b) : $g_t(f_i) = (1-\lambda_i r)(1-r)$

For i sufficiently large, we have $|\, x_i(t) \,| < \delta r/3$. It follows from Prop. 25.6 (taking $f_1 = x_1 = 0$) that there exists $f \in F_t(x_i) \cap B(0, k\,|\, x_i \,|)$. Then

$$\begin{aligned} g_t(f) &\leqslant g_t(0) + (\delta k)^{-1} |\, f \,| \leqslant (\delta k)^{-1} k\, |\, x_i \,|\\ &< (\delta k)^{-1} k\, \delta\, r = r < (1-r)^2 \;\; \text{(by (1))},\end{aligned}$$

so that the point $(f, 1)$ belongs to $F_t^{\,r}(x_i)$. We also find $\Lambda_t(f, 1) = 0$, since

$$g_t(v) - \lambda(1-r)(1-r) + r \leqslant r + (1-r)^2 = -1 + 4r - r^2 < 0,$$

because $r < 1/4$ by (1). We have in addition $\Lambda_t(\lambda_i f_i, \lambda_i) = \lambda_i r$, as a result of the equality characterizing Case 3(b). Substitution in (11) for this choice of f and λ now leads to

$$\begin{aligned}\langle p_i, \lambda_i f_i - f\rangle &\geqslant q_i(1-\lambda_i) + q_i \Lambda_t(\lambda_i f_i, \lambda_i) - \varepsilon_i(1+2M)\\ &= q_i(1-\lambda_i) + q_i r \lambda_i - \varepsilon_i(1+2M)\\ &\geqslant q_i r - \varepsilon_i(1+2M) > r/2 - \varepsilon_i(1+2M) \text{ (since } q_i > 1/2).\end{aligned}$$

When combined with the estimate $|\lambda_i f_i - f| \leqslant 2M$, this implies that, provided ε_i satisfies $\varepsilon_i(1+2M) \leqslant r/3$, we must have $|p_i| \geqslant r/(12M)$, confirming (19).

Let us note the following fact regarding the arc p we have produced:

25.10 Proposition. *The arc p satisfies*

$$p'(t) \in \operatorname{co}\{\omega : (\omega, p(t)) \in N^L_{G_t}(0,0)\}, \ t \in [a,b] \text{ a.e.} \tag{20}$$

Proof. This will be seen to be a consequence of (17). Because the limiting cone is generated from the proximal one by limits, it suffices to prove

$$(\omega, p, q) \in N^P_{G^r_t}(x, \lambda f, \lambda), \ g_t(f) < (1-r)^2/2 \implies (\omega, \lambda p) \in N^P_{G_t}(x,f).$$

Given data as described on the left of this putative implication, the definition of proximal normal (Def. 11.25) yields the existence of some $\sigma \geqslant 0$ such that:

$$\langle (\omega,p,q),(x'-x, \lambda' f' - \lambda f, \lambda' - \lambda)\rangle \leqslant \\ \sigma\{|x'-x|^2 + |\lambda' f' - \lambda f|^2 + |\lambda' - \lambda|^2\} \ \forall (x', \lambda' f', \lambda') \in G^r_t.$$

Let $f' \in F_t(x')$. If, in addition to this inclusion, f' is sufficiently close to f, then we have $g_t(f') < (1-r)^2$, so that the point $(\lambda f', \lambda)$ belongs to $F^r_t(x')$. The preceding inequality for this choice yields

$$\langle (\omega, \lambda p), (x'-x, f'-f)\rangle \leqslant \sigma\{|x'-x|^2 + |f'-f|^2\},$$

which holds therefore for all points $(x', f') \in G_t$ which are sufficiently close to (x,f). Thus $(\omega, \lambda p) \in N^P_{G_t}(x,f)$ (since proximal normals can be characterized either locally or globally), and the result follows. □

If in (16) we set $\eta = \gamma q$, and if we normalize by dividing across by $\eta + \|p\|$, we arrive at the following, which summarizes our progress to this point:

25.11 Proposition. *There exist an arc p and $\eta \geqslant 0$ with $\eta + \|p\| = 1$ that satisfy the transversality condition*

$$(p(a), -p(b)) \in \partial_L \eta\, \ell(0,0) + N^L_E(0,0), \tag{21}$$

the Euler inclusion

$$p'(t) \in \operatorname{co}\{\omega : (\omega, p(t)) \in N^L_{G_t}(0,0)\} \text{ a.e.} \tag{22}$$

and the following maximum condition for almost every t:

$$f \in F_t(0),\; g_t(f) \leqslant (1-r)^2 - 2r \implies \langle p, f \rangle \leqslant 0. \tag{23}$$

These conclusions are very close to those asserted by the theorem; their only defect is that the upper bound on $g_t(f)$ in (23) does not allow all points $f \in F_t(0) \cap R_t$ to be involved in the assertion of the maximum condition.

Letting r shrink to zero. For each r_i of a sequence decreasing to 0, we obtain an arc p_i and a scalar η_i satisfying the properties given in Prop. 25.11; these imply the differential inequality $|p_i'| \leqslant k|p_i|$ a.e. We may deploy familiar arguments involving Gronwall's lemma, together with Prop. 25.3, to extract a subsequence of p_i and η_i converging to limits that continue to satisfy (25.5 c) and (25.5 b).

We now prove that (25.5 d) holds as well. Let t be such that the hypotheses hold at t, as well as (23) for each p_i. Let $f \in F(t,0) \cap R_t$; then $g_t(f) < 1$. We wish to prove that $\langle p(t), f \rangle \leqslant 0$. But for all i sufficiently large, we have $g_t(f) < (1-r_i)^2 - 2r_i$, as well as $\langle p_i(t), f \rangle \leqslant 0$ (by (23)). The required conclusion is obtained by passing to the limit.

Removing the temporary hypotheses. We next remove the temporary hypothesis [TH2], while still assuming [TH1] (and that $x_* \equiv 0$). To do so, we consider the original problem in which R_t is replaced by $R_t \cap B^{\circ}(0,M)$, for any $M > \delta k$. Then all the hypotheses [H3] to [H5] as well as [TH1] are still valid, and [TH2] holds as well. We may therefore apply to this problem the case of the theorem that has been proved. In so doing we obtain all the assertions of Theorem 25.5, except that we have $R_t \cap B^{\circ}(0,M)$ instead of R_t in the maximum condition, and the nontriviality is expressed in the form $\eta + \|p\| = 1$. We proceed to let M go to $+\infty$ along a sequence M_i. By the usual sequential compactness and closure arguments (using notably Prop. 25.3), we extract a subsequence of the corresponding p_i and η_i converging to limits p and η which satisfy all the necessary conditions of Theorem 25.5.

There remains to remove temporary hypothesis [TH1]; we do so by time-rescaling, exactly as in the final step of the proof of Theorem 25.1. We omit the details.

The global case. We shall say that a sequence of multifunctions R_t^i of the type considered in Theorem 25.5 *diverges to* $\mathbb{R}^n$ (relative to x_*), and we write $R_t^i \to \mathbb{R}^n$, provided that for every $M > 0$ there exists i_M such that

$$i \geqslant i_M \implies R_t^i \supset B(x_*'(t), M) \text{ a.e.}$$

25.12 Corollary. *Let R_t^i be a sequence satisfying $R_t^i \to \mathbb{R}^n$ relative to x_*, where x_* satisfies the hypotheses of* Theorem 25.5 *for each radius R_t^i, with corresponding data ε^i, k^i, δ^i that may depend on i. Then all the conclusions of* Theorem 25.5 *hold for an arc p satisfying the following* global *maximum condition*: *for almost every t we have*

$$\langle p(t), v \rangle \leqslant \langle p(t), x_*'(t) \rangle \;\; \forall v \in F_t\big(x_*(t)\big).$$

Proof. The same sequential compactness and limiting arguments used repeatedly above are used to derive this corollary. For each i, the theorem (applied to x_*) yields η_i and p_i satisfying the transversality condition and the Euler inclusion, together with the maximum condition relative to the radius R_t^i. Normalization is applied to arrange $\eta_i + \|p_i\| = 1$.

Then the arcs p_i satisfy the differential inequality $|p_i'| \leqslant k^1 |p_i|$, where k^1 is the summable function corresponding via hypothesis [H4] to the first radius R_t^1. This allows the usual sequential compactness argument to be made (with the help of Prop. 25.3), leading to η and p that satisfy the transversality condition, the Euler inclusion, and $\eta + \|p\| = 1$. For almost every t, we have, for all i,

$$\max \big\{ \langle p_i(t), v \rangle : v \in F_t\big(x_*(t)\big) \cap R_t^i \big\} \leqslant \langle p_i(t), x_*'(t) \rangle .$$

Since p_i converges uniformly to p, and since $R_t^i \to \mathbb{R}^n$, a routine argument derives the global maximum condition. □

25.13 Example. We proceed to illustrate the utility of the radius introduced in Theorem 25.5 by deriving the necessary conditions for the general problem of Bolza, as given in Theorem 18.13.

We first augment the state x by an additional coordinate y, in order to absorb the integral cost into the dynamics. This is done by defining

$$F_t(x,y) = \big\{ [v, \Lambda(t,x,v) + \rho] : v \in \mathbb{R}^n,\ \rho \geqslant 0 \big\}$$
$$\ell_+(x_0, y_0, x_1, y_1) = \ell(x_0, x_1) + y_1,\ E = \big\{ (x_0, y_0, x_1, y_1) : (x_0, x_1) \in S,\ y_0 = 0 \big\}$$
$$y_*(t) = \int_a^t \Lambda\big(s, x_*(s), x_*'(s)\big)\, ds .$$

It is a notational exercise to see that the augmented arc (x_*, y_*) provides a local minimum for the corresponding version of the problem (DI) considered in Theorem 25.5; that is, a minimum relative to trajectories (x,y) for which $\|x - x_*\| < \varepsilon$.

We denote $\Lambda(t, x_*(t), x_*'(t))$ by $\Lambda_*(t)$ and $x_*'(t)$ by $v_*(t)$. For $M > 0$, we set

$$\rho_M(t) = M \big\{ 1 + |(v_*(t), \Lambda_*(t))| + d(t) \big\},$$

and we define the radius $R_M(t)$ to be the open ball of radius $\rho_M(t)$ around the point $(v_*(t), \Lambda_*(t))$. We proceed to verify the hypotheses [H3][H4][H5] for these data. The point of using a radius here is that the bounded slope condition holds for each radius R_M, but not globally (that is, for the radius $R(t) = \mathbb{R}^n \times \mathbb{R}$); it is a situation in which Cor. 25.12 will serve.

That [H3] holds is easy to see: the requisite measurability of the multifunction

$$t \mapsto \big\{ \big(x, v, \Lambda(t,x,v) + \rho\big) : \rho \geqslant 0 \big\}$$

follows from the fact that it is closed-valued (as follows from the lower semicontinuity of Λ with respect to (x,v)), and that its graph is LB measurable (because of the LB measurability of Λ). Consider [H4], the bounded slope condition. We suppress the t variable for ease of notation. Let

$$(\alpha_+, \beta_+) \in N_G^P\big(x,y,v,w\big), \text{ where } |x-x_*| < \varepsilon, \ (v,w) \in F(x,y) \cap R_M(t).$$

Since F does not depend on y, and since G is essentially the epigraph of $\Lambda(\cdot,\cdot)$, it follows that α_+ is of the form $(\alpha,0)$ and that β_+ is of the form $(\beta,-\lambda)$ for some $\lambda \geqslant 0$. Suppose first $\lambda > 0$. Then $w = \Lambda(x,v)$ (see Exer. 11.30), and we have

$$\big(\alpha/\lambda, \beta/\lambda\big) \in \partial_P \Lambda(x,v).$$

Invoking the Tonelli-Morrey condition 18.11, where S is any bounded set containing an ε-neighborhood of $\{x_*(t) : t \in [a,b]\}$, we deduce from this

$$|\alpha|/(\lambda + |\beta|) \leqslant c\big(|v| + |\Lambda(x,v)|\big) + d,$$

which leads to

$$\begin{aligned} |\alpha_+| = |\alpha| &\leqslant 4\big\{c\,\big|(v,\Lambda(x,v))\big| + d\big\}\,|(\beta,\lambda)| \\ &\leqslant 4\big\{c\big(|(v_*,\Lambda_*)| + \rho_M\big) + d\big\}\,|\beta_+| \leqslant \bar{c}\,(1+M)\big(1 + |(v_*,\Lambda_*)| + d\big)\,|\beta_+|, \end{aligned}$$

for a constant $\bar{c}$ not depending on M or t. This confirms the bounded slope condition of radius R_M, with the summable function

$$k_M = \bar{c}\,(1+M)\big\{1 + \big|\big(v_*(t), \Lambda_*(t)\big)\big| + d(t)\big\}.$$

If $\lambda = 0$, then a straightforward approximation argument based upon Theorem 11.31 and the analysis above of the case $\lambda > 0$ gives rise to the same inequality. Since $\rho_M/k_M = M/[\bar{c}\,(1+M)]$, we see that [H5] is satisfied as well.

Because the radius multifunctions R_M diverge to $\mathbb{R}^{n+1}$ in the sense of Cor. 25.12, we deduce the existence of η and an augmented costate (denoted by (p,q)) satisfying the conclusions of Theorem 25.5 for the augmented problem. The Euler inclusion implies that q is constant, and then the transversality condition yields $q = -\eta$. If $\eta = 0$, the maximum condition asserts that, for almost every t,

$$\langle p(t), v\rangle \leqslant \langle p(t), x_*'(t)\rangle \ \ \forall\, v \in \mathbb{R}^n.$$

But then $p = 0$ and nontriviality is violated. Thus, $\eta = 1$. It now follows that p satisfies the conclusions of Theorem 18.13.

There remains the autonomous case to consider, with its extra conclusion. The now familiar time-rescaling device will be used. We modify the augmented problem above by taking

$$F_+(x,y,z) = \big\{[\,wv, w\Lambda(x,v) + \rho, w\,] : v \in \mathbb{R}^n, \ \rho \geqslant 0, \ w \in [1-\delta, 1+\delta]\,\big\},$$

where $\delta > 0$ is sufficiently small in a sense specified below. The additional state coordinate z is subject to prescribed boundary conditions $z(a) = a,\ z(b) = b$, and we set $z_*(t) = t$. Then any trajectory (x, y, z) of F_+ induces a trajectory $(\tilde{x}, \tilde{y})$ of F, where we define

$$(\tilde{x}, \tilde{y})(\tau) = (x, y)\big(t(\tau)\big), \text{ with } \tau(t) = a + \int_a^t w(s)\, ds.$$

If δ is suitably small, then $\tilde{x}_*$ is uniformly within ε of x_*, and it follows that (x_*, y_*, z_*) is a local minimum for the new problem.

The arguments given above in the non autonomous case adapt without difficulty to show the applicability of Cor. 25.12 to the problem. Then the necessary conditions are obtained, for a costate that we denote by (p, q, r). As before, we find $q = -1$; we see that r is constant. There follow the same conclusions as previously, but the presence of the additional control variable w gives rise to an extra conclusion: almost everywhere, the maximum over $[\,1-\delta, 1+\delta\,]$ of the expression

$$w\big\{\,\langle\, p(t), v_*(t)\rangle - \Lambda_*(t) + r\,\big\}$$

occurs at the interior point $w_* = 1$. Thus, the quantity in braces vanishes, which implies the Erdmann condition (with $h = -r$). □

25.4 The multiplier rule and mixed constraints

We consider in this section an optimal control problem with standard cost and dynamics, but in which the state x and control u are subject to additional joint, or *mixed constraints* of the type $\varphi(t, x(t), u(t)) \in \Phi \subset \mathbb{R}^k$. The presence of such constraints has long been known to constitute a challenge as regards the derivation of appropriate necessary conditions of maximum principle type. Specifically, then, the problem is the following:

$$\left\{\begin{array}{lll} \text{Minimize} & J(x,u) = \ell\big(x(a), x(b)\big) + \displaystyle\int_a^b \Lambda\big(t, x(t), u(t)\big)\, dt & \\ \text{subject to} & x'(t) = f\big(t, x(t), u(t)\big),\ u(t) \in U,\ \ t \in [a,b] \text{ a.e.} & \textbf{(MC)} \\ & \varphi\big(t, x(t), u(t)\big) \in \Phi,\ \ t \in [a,b] \text{ a.e.} & \\ & \big(x(a), x(b)\big) \in E. & \end{array}\right.$$

The functions ℓ, f, Λ, and φ satisfy the classical regularity hypotheses 22.1. The sets U, E, and Φ are closed.[6]

[6] A more general approach to the results of this section is given in Clarke and de Pinho [17].

Let (x_*, u_*) be a local minimizer for (MC), where we assume that u_* is bounded. The crucial hypothesis made on the nature of the mixed constraints is the following: for every $t \in [a,b]$, we have

$$u \in U,\ \varphi\big(t, x_*(t), u\big) \in \Phi,\ \lambda \in N^L_\Phi\big(\varphi(t, x_*(t), u)\big)$$
$$0 \in D_u\langle \lambda, \varphi\rangle\big(t, x_*(t), u\big) + N^L_U(u) \implies \lambda = 0. \quad (1)$$

We remark that this is a natural constraint qualification (or nondegeneracy condition) of the type encountered in Theorem 11.38. When $\Phi = \{0\}$ and $U = \mathbb{R}^m$, (1) is equivalent to requiring that $D_u\,\varphi(t, x_*(t), u)$ have rank k at points $u \in U$ for which $\varphi(t, x_*(t), u) = 0$. When $\Phi = \mathbb{R}^k_-$, it is equivalent to the corresponding positive linear independence condition.

It is convenient to define an augmented Hamiltonian

$$\begin{aligned} H^\eta_\varphi(t, x, p, u, \lambda) &= H^\eta(t, x, p, u) - \langle \lambda, \varphi\rangle(t, x, u) \\ &= \langle p, f(t, x, u)\rangle - \eta\,\Lambda(t, x, u) - \langle \lambda, \varphi\rangle(t, x, u). \end{aligned}$$

25.14 Theorem. *Under the hypotheses above, there exists an arc $p : [a,b] \to \mathbb{R}^n$, a scalar η equal to 0 or 1, and a bounded measurable function $\lambda : [a,b] \to \mathbb{R}^k$ with*

$$\lambda(t) \in N^C_\Phi\big(\varphi(t, x_*(t), u_*(t))\big) \text{ a.e.} \quad (2)$$

satisfying the **nontriviality condition**

$$\big(\eta, p(t)\big) \neq 0\ \ \forall t \in [a,b], \quad (3)$$

the **transversality condition**

$$\big(p(a), -p(b)\big) \in \eta \nabla \ell\big(x_*(a), x_*(b)\big) + N^L_E\big(x_*(a), x_*(b)\big), \quad (4)$$

the **adjoint equation** *for almost every t:*

$$-p'(t) = D_x H^\eta_\varphi\big(t, x_*(t), p(t), u_*(t), \lambda(t)\big) \quad (5)$$

as well as, for almost every t, the **maximum condition**

$$u \in U,\ \varphi\big(t, x_*(t), u\big) \in \Phi \implies$$
$$H^\eta\big(t, x_*(t), p(t), u\big) \leqslant H^\eta\big(t, x_*(t), p(t), u_*(t)\big) \quad (6)$$

and the **stationarity condition**

$$0 \in D_u H^\eta_\varphi\big(t, x_*(t), p(t), u_*(t), \lambda(t)\big) - N^C_U\big(u_*(t)\big). \quad (7)$$

Remark. It is not hard to see that stationarity (7) is a natural necessary condition corresponding to the constrained maximization in (6), but an extra assertion is being

made: the *same* normal vector λ in N^C_Φ occurs here as in the adjoint equation. Note that when $\Phi = \mathbb{R}^m$, then $\lambda \equiv 0$, and we recover precisely the conclusions of the usual maximum principle.

Proof. The proof is based on appealing to Theorem 25.5 or, more precisely, Cor. 25.12. Fix any $r > 0$. We show first that condition (1) holds locally around x_*, if u is constrained to the ball $B(u_*(t), r)$.

Lemma 1. *There exists* $\varepsilon_1 > 0$ *and* M *such that*

$$t \in [a,b],\ |x - x_*(t)| \leqslant \varepsilon_1,\ u \in U,\ |u - u_*(t)| \leqslant r,\ \varphi\big(t,x,u\big) \in \Phi,$$
$$\lambda \in N^L_\Phi\big(\varphi(t,x,u)\big),\ \psi \in D_u\langle \lambda, \varphi\rangle\big(t,x,u\big) + N^L_U(u) \implies |\lambda| \leqslant K|\psi|.$$

Proof. We argue by contradiction. If the lemma is false, there exist sequences

$$t_i \in [a,b],\ \varepsilon_i \downarrow 0,\ x_i \in B\big(x_*(t_i), \varepsilon_i\big),\ u_i \in U \cap B\big(u_*(t_i), r\big),\ \lambda_i \in N^L_\Phi\big(\varphi(t_i, x_i, u_i)\big)$$

such that, for certain ψ_i, we have

$$\psi_i \in D_u\langle \lambda_i, \varphi\rangle\big(t_i, x_i, u_i\big) + N^L_U(u_i),\ |\lambda_i| > n_i|\psi_i|,\ n_i \to \infty.$$

We may normalize to have $|\lambda_i| = 1$, and then take subsequences to suppose that

$$\lambda_i \to \lambda_0,\ t_i \to t_0,\ u_i \to u_0 \in U.$$

It follows that $\psi_i \to 0$, $x_i \to x_*(t_0)$ and $\varphi(t_0, x_*(t_0), u_0) \in \Phi$. In the limit, we discover

$$0 \in D_u\langle \lambda_0, \varphi\rangle\big(t_0, x_*(t_0), u_0\big) + N^L_U(u_0),$$

which contradicts (1) (since λ_0 is a unit vector) and proves the lemma. □

By the usual device of absorbing the running cost into the dynamics (see for example page 518), there is no loss of generality in taking $\Lambda \equiv 0$, as we do henceforth.

We now proceed to augment the state x with another component $y \in \mathbb{R}^m$; thus, the new state is (x, y). In these terms, we define the ingredients of an augmented problem of the form (DI) (see p. 521):

$$F_t(x,y) = \big\{ (v,u) : v - f(t,x,u) = 0,\ u \in U,\ \varphi(t,x,u) \in \Phi \big\}$$
$$\ell_+(x_0, y_0, x_1, y_1) = \ell(x_0, x_1),\ E_+ = \big\{ (x_0, y_0, x_1, y_1) : (x_0, x_1) \in E,\ y_0 = 0 \big\}.$$

It is a notational exercise to check that (x_*, y_*) is a local minimizer for the resulting version of (DI), where (note that u_* is bounded, and therefore integrable)

$$y_*(t) := \int_a^t u_*(s)\,ds.$$

We wish to apply Theorem 25.5, with radius $R_t = \mathbb{R}^n \times B^\circ(u_*(t), r)$. To this end, let us verify that the hypotheses of Theorem 25.5 are satisfied. Since [H3] is easily seen to hold, it suffices to prove that the bounded slope condition [H4] holds for some $\varepsilon > 0$, and with a *constant* k (for then [H5] certainly holds for δ sufficiently small).

We argue by contradiction. If this fails, there exist sequences

$$t_i \in [a,b],\ \varepsilon_i \downarrow 0,\ x_i \in B\big(x_*(t_i), \varepsilon_i\big),\ y_i \in B\big(y_*(t_i), \varepsilon_i\big),\ u_i \in U \cap B\big(u_*(t_i), r\big)$$

and points

$$\big(\alpha_i, c_i, \beta_i, \zeta_i\big) \in N^P_{G(t_i)}\big(x_i, y_i, f(t_i, x_i, u_i), u_i\big) \tag{8}$$

such that

$$1 = |(\alpha_i, c_i)| = |\alpha_i| > n_i \big|(\beta_i, \zeta_i)\big|,\ \ n_i \to \infty. \tag{9}$$

In writing this, we have normalized in order to have $|(\alpha_i, c_i)| = 1$, and then used the fact that $c_i = 0$, since F does not depend on y. We have also written $G(t)$ for the graph of F_t. We may omit y, and regard $G(t_i)$ as being the set

$$\big\{ (x,v,u) : \big(v - f(t_i, x, u), u, \varphi(t_i, x, u)\big) \in \{0\} \times U \times \Phi \big\}.$$

Lemma 2. *For i sufficiently large, there exist $(\gamma_i, \tau_i, \lambda_i)$ with $\lambda_i \in N^L_\Phi(\varphi(t_i, x_i, u_i))$ and $\tau_i \in N^L_U(u_i)$ such that*

$$\begin{aligned}\big[\alpha_i, \beta_i, \zeta_i\big] &= D_{x,v,u}\big\{\langle \gamma_i, v - f(x,u)\rangle + \langle \tau_i, u\rangle + \langle \lambda_i, \varphi\rangle\big\}\big(x_i, f(t_i, x_i, u_i), u\big)\\ &= \big[D_x\big(\langle \lambda_i, \varphi\rangle - \langle \gamma_i, f\rangle\big)(x_i, u_i),\ \gamma_i,\ D_u\big(\langle \lambda_i, \varphi\rangle - \langle \gamma_i, f\rangle\big)(x_i, u_i) + \tau_i\big].\end{aligned}$$

Proof. It is a matter of showing that Theorem 11.38 applies. Let i be large enough to ensure that $x := x_i$ satisfies $|x - x_*(t_i)| < \varepsilon_1$, where ε_1 is provided by Lemma 1. Suppose that, for some $\lambda \in N^L_\Phi(x,u)$ and $\tau \in N^L_U(u)$, we have

$$[0,0,0] = D_{x,v,u}\big\{\langle \gamma, v - f(x,u)\rangle + \langle \tau, u\rangle + \langle \lambda, \varphi\rangle\big\}(x,v,u).$$

Then we find $\gamma = 0$, so that $0 = D_u\langle \lambda, \varphi\rangle(x,u) + \tau$, whence $\lambda = 0$ by Lemma 1. This confirms the applicability of Theorem 11.38, which yields the stated characterization of $[\alpha_i, \beta_i, \zeta_i]$ and proves Lemma 2. □

We now return to the situation described by (8) and (9). Observe that $\beta_i \to 0$ and $\zeta_i \to 0$. Define a compact set A by

$$A = \big\{ (t,x,u) : t \in [a,b],\ |x - x_*(t)| \leqslant \varepsilon,\ u \in U,\ |u - u_*(t)| \leqslant r \big\}.$$

Letting K be a Lipschitz constant for f and φ on A, we see, with the aid of Lemma 2, that $\beta_i = \gamma_i$, and we calculate

$$1 = |\alpha_i| \leqslant K|\lambda_i| + K|\gamma_i| = K\big(|\lambda_i| + |\beta_i|\big).$$

It follows that, for some $\rho > 0$, we have $|\lambda_i| \geqslant \rho$ for all large i. We may suppose that

$$t_i \to t\,,\ u_i \to u \quad \text{and} \quad \lambda_i/|\lambda_i| \to \lambda \neq 0\,.$$

Then $x_i \to x_*(t)$ and $\lambda \in N^L_\Phi(\varphi(t, x_*(t), u))$. Dividing the expression for ζ_i by $|\lambda_i|$ and passing to the limit, we obtain

$$0 \in D_u\langle \lambda, \varphi \rangle\big(t, x_*(t), u\big) + N^L_U(u)\,,$$

contradicting (1), and confirming the bounded slope condition [H4].

Thus, the hypotheses of Theorem 25.5 hold, for the augmented problem defined in terms of F_t, ℓ_+, and E_+, for *any* radius of the form

$$R_t = \mathbb{R}^n \times B(u_*(t), r).$$

We may therefore invoke Cor. 25.12, which yields a scalar η and an arc that we denote by (p, q). From the Euler inclusion and transversality condition, we find that q is identically zero. It follows that $(\eta, p(t)) \neq 0\ \forall t$ (the required nontriviality), and that the maximum condition (6) and the transversality condition (4) hold. The Euler inclusion affirms that (for almost every t, which we suppress)

$$p' \in \operatorname{co}\big\{\, \omega : (\omega, p, 0) \in N^L_S\big(x_*, x_*', u_*\big)\big\},$$

where

$$S := \big\{\, (x,v,u) : v - f(x,u) = 0,\ u \in U,\ \varphi(x,u) \in \Phi \,\big\}.$$

For any such ω as described above, we may invoke Theorem 11.38 once again in order to deduce the existence of

$$(\gamma, \tau, \lambda) \ \text{ with } \ \lambda \in N^L_\Phi(\varphi(x_*, u_*)),\ \tau \in N^L_U(u_*)$$

for which the following holds:

$$(\omega, p, 0) = D_{x,v,u}\big\{\, \langle \gamma, v - f(x,u) \rangle + \langle \tau, u \rangle + \langle \lambda, \varphi \rangle \,\big\},$$

where the derivative is evaluated at (x_*, x_*', u_*). It follows that $\gamma = p$, whence

$$\omega = -D_x\big\{\langle p, f \rangle - \langle \lambda, \varphi \rangle\big\}(x_*, p, u_*) = -D_x H_\varphi\big(x_*, p, u_*, \lambda\big)$$
$$0 \in D_u\big\{\langle p, f \rangle - \langle \lambda, \varphi \rangle\big\}(x_*, u_*) - N^L_U(u_*) = D_u H_\varphi\big(x_*, p, u_*, \lambda\big) - N^L_U(u_*).$$

Since $|D_u\langle p, f \rangle(x_*, u_*)| \leqslant K\|p\|$, the last equation, in light of Lemma 1, yields $|\lambda| \leqslant MK\|p\|$. Since p' is a convex combination of points ω as above, and since $\operatorname{co} N^L = N^C$ (Theorem 11.36), we deduce the adjoint equation (5), for some new λ satisfying (2) and (7), and which continues to be bounded by $MK\|p\|$.

That λ can be chosen to be a measurable function is an exercise in measurable selections. This completes the proof of Theorem 25.14. □

Proof of two multiplier rules. We proceed below to derive from Theorem 25.14 two multiplier rules encountered in Chapter 17.

25.15 Corollary. Theorem 17.4 *holds.*

Proof. This is an immediate consequence of Theorem 25.14. We simply specialize to the case of the trivial dynamics $x' = f(t,x,u) = u$, with the control set $U = \mathbb{R}^n$, and we take $E = \{(A,B)\}$ and $\ell \equiv 0$. □

There remains one more theorem from Chapter 17 to prove, namely the classical multiplier rule for the problem of Lagrange in the calculus of variations.

25.16 Corollary. Theorem 17.1 *holds.*

Proof. We define

$$F_t(x,y) = \big\{ \big(v, \Lambda(t,x,v)\big) : \varphi(t,x,v) = 0 \big\},\ y_*(t) = \int_a^t \Lambda\big(s, x_*(s), x_*'(s)\big)\,ds$$
$$\ell\big(x_0,y_0,x_1,y_1\big) = y_1,\ E = \big\{ \big(x_0,y_0,x_1,y_1\big) : x_0 = A,\ y_0 = 0,\ x_1 = B \big\}.$$

It follows from the fact that x_* is a weak local minimizer for the original problem that the arc (x_*, y_*) provides a local minimizer of radius R for the problem (DI) corresponding to these data, for a certain radius multifunction having the form

$$R_t = B^{\circ}(x_*'(t), \rho) \times \mathbb{R}, \tag{10}$$

for some $\rho > 0$. The next result verifies that F satisfies the bounded slope condition near (x_*, y_*). We denote by $G(t)$ the graph of $F_t(\cdot,\cdot)$.

Lemma. *There exists* $\delta > 0$ *and* $k \geqslant 0$ *such that*

$$t \in [a,b],\ |x - x_*(t)| \leqslant \delta,\ \varphi(t,x,v) = 0,\ v \in B(x_*'(\tau+),\delta) \cup B(x_*'(\tau-),\delta),$$
$$(\alpha,\theta,\beta,\gamma) \in N^P_{G(t)}(x,y,v,w) \implies \theta = 0,\ |\alpha| \leqslant k|(\beta,\gamma)|.$$

Proof. That $\theta = 0$ is a consequence of the fact that F does not depend on y. Suppose now that the lemma is false. Then (suppressing the y variable) there exists a sequence $t_i \to \tau \in [a,b]$ and sequences

$$\big(\alpha_i, \beta_i, \gamma_i\big) \in N^P_{G(t_i)}(x_i, v_i, w_i),\ |\alpha_i| > n_i|(\beta_i,\gamma_i)|$$

where $n_i \to \infty$ and

$$(x_i, v_i, w_i) \to \big(x_*(\tau), x_*'(\tau), y_*'(\tau)\big).$$

(One interprets this with the appropriate one-sided derivative if x_* has a corner at τ.) By scaling, we may take $|\alpha_i| = 1$, so that $(\beta_i,\gamma_i) \to 0$. The set $G(t_i)$ may be expressed as a level set:

$$G(t_i) = \{ (x,v,w) : \varphi(t_i,x,v) = 0,\ w - \Lambda(t_i,x,v) = 0 \}.$$

According to Theorem 11.38, and because $D_v \varphi(x_i, v_i, w_i)$ has rank k for i sufficiently large, there exist λ_i and r_i such that

$$\alpha_i = D_x \{ \langle \lambda_i, \varphi \rangle - r_i \Lambda \}(t_i, x_i, v_i),\ (\beta_i, \gamma_i) = \big(D_v \{ \langle \lambda_i, \varphi \rangle - r_i \Lambda \}(t_i, x_i, v_i), r_i \big).$$

Thus $r_i = \gamma_i \to 0$. Since $|\alpha_i| = 1$, there exists $\varepsilon > 0$ such that $|\lambda_i| \geqslant \varepsilon$ for all i sufficiently large. Then $\lambda_i / |\lambda_i|$ converges to a unit vector $\bar\lambda$ for a subsequence, and from $(\beta_i, \gamma_i)/\lambda_i \to 0$ we deduce $D_v \langle \bar\lambda, \varphi \rangle (\tau, x_*(\tau), x_*'(\tau)) = 0$. This contradicts the rank hypothesis, and proves the lemma.

We are now authorized to apply Theorem 25.5, with a radius of the type given by (10). The same arguments used in the proof of Theorem 25.14 may be used to interpret the resulting Euler inclusion, in order to obtain (5), which, in the current setting, gives the required conclusion. □

Remark. The proof shows that (in view of (7)) we can go beyond classical results, in also asserting the stationarity condition

$$0 = D_u H_\varphi^\eta \big(t, x_*(t), p(t), u_*(t), \lambda(t)\big)$$

along with the Euler equation. In fact, we obtain somewhat more, a *local* Weierstrass condition: for some $\delta > 0$, for every t, we have

$$|v - x_*'(t)| < \delta,\ \varphi\big(t, x_*(t), v\big) = 0 \implies$$
$$\eta\, \Lambda\big(t, x_*(t), v\big) - \eta\, \Lambda\big(t, x_*(t), x_*'(t)\big) \geqslant \langle p(t), v - x_*'(t) \rangle,$$

with $x_*'(t)$ interpreted as either one-sided derivative if x_* has a corner at t.

Sufficiency of the multiplier rule. We proceed to formalize below a type of result that is now familiar to the reader: the necessary conditions provided by Theorem 25.14 for the problem (MC) (p. 535), in their normal form, are also *sufficient* when the underlying context is linear/convex.

We continue to assume that the functions ℓ, f, Λ, and φ satisfy the classical regularity hypotheses 22.1, and that the sets U, E, and Φ are closed. In addition, however, we now assume that:

(a) ℓ, E, U, and Φ are convex;

(b) Λ is convex in (x,u);

(c) f is of the form $Ax + Bu$ for certain matrices A and B;

(d) φ is of the form $Cx + Du$ for certain matrices C and D.

The proof of the following is the subject of Exer. 26.16.

25.17 Theorem. *Let the data of the problem* (MC) *satisfy the hypotheses above, and let* (x_*, u_*) *be an admissible process. Suppose that there exists an arc* p *and a bounded measurable function* λ *with*

$$\lambda(t) \in N_\Phi^C\big(\varphi(t, x_*(t), u_*(t))\big) \text{ a.e.}$$

that satisfy the transversality condition, adjoint equation, and stationarity condition of Theorem 25.14*, with* $\eta = 1$*. Then the process* (x_*, u_*) *is optimal for* (MC).

25.18 Example. We now illustrate the use of Theorem 25.17 in a simple case, that in which the problem considered in Example 22.9 is augmented by a mixed inequality constraint $x - u \leqslant 1/2$. Thus the problem is the following:

$$\begin{cases} \text{Minimize} & J(x,u) = \int_0^3 \big(x(t) + u(t)^2/2\big)\,dt \\ \text{subject to} & x'(t) = u(t) \in [-1,1] \text{ a.e.} \\ & x(t) - u(t) - 1/2 \leqslant 0 \text{ a.e.} \\ & x(0) = 0. \end{cases}$$

It is an exercise in the direct method to prove that a solution (x_*, u_*) exists. In the absence of the mixed constraint, the solution had been found to be

$$u_*(t) = \begin{cases} -1 & \text{if } 0 \leqslant t < 2 \\ t-3 & \text{if } 2 \leqslant t \leqslant 3 \end{cases} \qquad x_*(t) = \begin{cases} -t & \text{if } 0 \leqslant t < 2 \\ (t-3)^2/2 - 5/2 & \text{if } 2 \leqslant t \leqslant 3. \end{cases}$$

If this process happens to satisfy the mixed constraint $x - u \leqslant 1/2$, then of course it remains optimal for the new problem. However, we see that the constraint is violated on the interval $[0, 1/2)$, where we have $x - u > 1/2$.

In an effort to track the original solution as closely as possible, we set $x - u = 1/2$ initially. This gives rise to

$$x(t) = (1 - e^t)/2, \quad u(t) = -e^t/2.$$

This is feasible (that is, $u(t)$ lies in $[-1, 1]$) only until $t = \ln 2$, at which point we may switch to the control values that had been used before.

We arrive, then, at the following educated guess concerning the optimal control for the problem above:

$$u_*(t) = \begin{cases} -e^t/2 & \text{if } 0 \leqslant t < \ln 2 \\ -1 & \text{if } \ln 2 \leqslant t < 2 \\ t-3 & \text{if } 2 \leqslant t \leqslant 3. \end{cases}$$

This control corresponds to the following state trajectory x_*:

$$x_*(t) = \begin{cases} (1-e^t)/2 & \text{if } 0 \leqslant t < \ln 2 \\ \ln 2 - t - 1/2 & \text{if } \ln 2 \leqslant t < 2 \\ (t-3)^2/2 - 3 + \ln 2 & \text{if } 2 \leqslant t \leqslant 3. \end{cases}$$

We proceed to show that the hypotheses of Theorem 25.17 are satisfied by the process above, with $\eta = 1$, for a suitable costate p and multiplier λ. The mixed constraint corresponds to the data

$$\varphi(t,x,u) = x - u - 1/2, \quad \Phi = (-\infty, 0].$$

It is clear that the linear/convex structure postulated by Theorem 25.17 is present. We see that (for the process given above) the mixed constraint is saturated on the interval $[0, \ln 2]$, and slack otherwise, so that (2) of Theorem 25.14 requires that the multiplier λ satisfy

$$\lambda(t) \geqslant 0, \; t \in (0, \ln 2) \text{ a.e.}, \quad \lambda(t) = 0, \; t \in (\ln 2, 3) \text{ a.e.}$$

The Hamiltonian H_φ of the problem is the function

$$pu - x - u^2/2 - \lambda(x - u - 1/2).$$

It follows from the adjoint equation that $p' = 1$ on $(\ln 2, 3)$. Transversality (see (4)) yields $p(3) = 0$, so that $p(t) = t - 3$ on $[\ln 2, 3]$.

It is an easy matter to verify the stationarity condition (7)

$$0 \in D_u H_\varphi\big(t, x_*(t), p(t), u_*(t), 0\big) - N_U^C\big(u_*(t)\big)$$

on the interval $(\ln 2, 3)$. In the interval $(0, \ln 2)$, we have $u_*(t) \in \operatorname{int} U$, so that the stationarity condition (7) reduces to $p(t) - u_*(t) + \lambda(t) = 0$. This relation, combined with the adjoint equation $-p' = -1 - \lambda$ (see (5)), provides a differential equation for p whose solution is given by

$$p(t) = e^{-t}\big(2\ln 2 - 7\big) + 1 - e^t/4, \; t \in (0, \ln 2).$$

(We have imposed $p(\ln 2) = \ln 2 - 3$ in solving the equation, in order to preserve the continuity of p.) Then we find

$$\lambda(t) = e^{-t}\big(7 - 2\ln 2\big) - e^t/4 - 1, \; t \in (0, \ln 2).$$

We check without difficulty that λ is nonnegative on $(0, \ln 2)$, as required.

To sum up: we have exhibited a costate p and a multiplier λ that satisfy (for $\eta = 1$) the requirements of Theorem 25.17. It follows that the proposed process is optimal.
$\square$

Chapter 26

Additional exercises for Part IV

26.1 Exercise. Optimal control problems which arise in economics or finance often feature a *discount rate* δ. This parameter is used to express the present value of future revenues or expenses. We introduce this consideration in the problem described on p. 445, by modifying the cost functional (but nothing else) as follows:

$$J(x,u) = \int_0^T e^{-\delta t}\big(x(t)+u(t)\big)\,dt\,.$$

For small $\delta > 0$, it can be shown that the optimal process has the same general turnpike nature as before, but with a *different* turnpike value (depending on δ) for the state. Determine that value. □

26.2 Exercise. Suppose that in the problem described on p. 445, the cost functional (and nothing else) is modified as follows:

$$J(x,u) = x(T) + \int_0^T u(t)\,dt\,.$$

(a) Prove that the problem admits a solution.

(b) Prove that if $T > \ln 2$, the unique optimal control is piecewise continuous; more precisely, that it equals 0 up to a certain time $\tau \in (0,T)$, and then equals 3 thereafter. (Thus, the solution is of bang-bang rather than turnpike type.)

(c) Determine the switching time τ. □

26.3 Exercise. Consider the soft landing problem, Example 22.14, but suppose now that the mass of the lander decreases with time, due to fuel consumption. Specifically, we postulate the dynamics

$$m(t)\,x''(t) = u(t) \in [-1,+1]\,,$$

F. Clarke, *Functional Analysis, Calculus of Variations and Optimal Control*, Graduate Texts in Mathematics 264, DOI 10.1007/978-1-4471-4820-3_26,

where m is a positive-valued locally Lipschitz function. As before, the goal is to find a control that achieves $x(\tau) = x'(\tau) = 0$ in minimal time τ. Assuming that an optimal control exists, prove that, as in the original case $m \equiv 1$, it is bang-bang with at most one switch. □

26.4 Exercise. We consider the soft landing problem (Example 22.14) in which the cost is modified as follows:

$$J(x,u) = \int_0^\tau \big(|u(t)| + k\big)\,dt.$$

The term $|u|$ is a natural one to reflect fuel expenditure, and the parameter $k > 0$ allows us to modulate in varying proportions the importance assigned to travel time and fuel cost.

(a) Prove that a solution of the problem exists.

(b) Show that the optimal control is piecewise constant with at most two switches, and takes the values $1, 0, -1$, either in that order, or else the opposite. □

26.5 Exercise. A ship moves in the (x,y)-plane with maximal speed $V > 0$, and navigates subject to a current whose velocity vector

$$c(x,y) = \big(c_x(x,y), c_y(x,y)\big)$$

depends on position. The goal is to steer from a prescribed initial condition to a given target set E in minimal time. We obtain the following formulation of the problem:

$$\begin{cases} \text{Minimize} & J(x,u) = \tau = \displaystyle\int_0^\tau 1\,dt \\ \text{subject to} & \tau \geqslant 0 \\ & \big(x',y'\big) = (v,w) + c(x,y),\ t \in [0,\tau] \text{ a.e.} \\ & |(v,w)| \leqslant V,\ t \in [0,\tau] \text{ a.e.} \\ & \big(x(0),y(0)\big) = (x_0,y_0),\ \big(x(\tau),y(\tau)\big) \in E. \end{cases}$$

Assumptions: The function c is locally Lipschitz and has linear growth ; the target set E is closed, does not contain the initial point, but can be reached in finite time ; the boat is able to remain in the target after arrival:

$$|c(x,y)| < V \quad \text{for } (x,y) \in E.$$

(a) Prove that the problem admits an optimal control (v_*, w_*) on an interval $[0, \tau_*]$.

(b) Invoke necessary conditions to deduce that $|(v_*(t), w_*(t))| = V$ a.e., and that (v_*, w_*) is continuous (has a continuous representative).

(c) If $c = 0$ on ∂E, show that the terminal velocity of the optimal trajectory is an inward normal vector to the set E.

We denote by u_* the optimal steering angle defined by

$$(v_*(t), w_*(t)) = (V\cos u_*(t), V\sin u_*(t)).$$

Then u_* may be taken to be continuous, in view of the above.

(d) Suppose now that $c(x,y) = (c_0(y), 0)$; thus, the current depends only upon y and has horizontal effect. If the function $c_0(\cdot)$ is affine, and if we exclude the cases in which u_* is either identically $\pi/2$ or identically $-\pi/2$, show that for certain constants $\alpha \neq 0$ and β, we have

$$u_*(t) = \arctan(\alpha t + \beta) \;\; \forall t \in [0, \tau_*].$$

Thus u_* is either constant or strictly monotone. Based on these facts, optimal trajectories can be calculated; a typical one (for $E = \{0\}$ and $c_0(y) = -y$) is shown in the figure below.

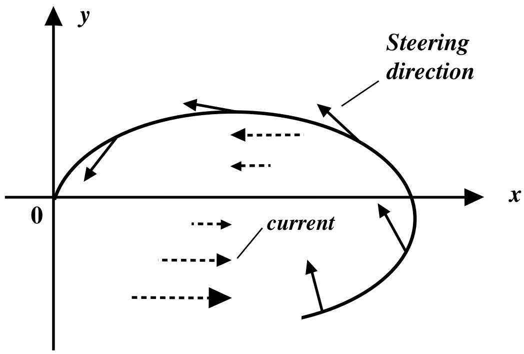


Fig. 26.1
Steering to the origin

(e) We now take $c(x,y) = (|y|, 0)$. Show that, in contrast to the previous case, it is now possible for an optimal trajectory to move along the x-axis. □

26.6 Exercise. Let $\theta : [0,1] \to \mathbb{R}$ be given by $\theta(t) = 2t - 1$. We seek the arc x mapping $[0,1]$ to $\mathbb{R}$ that is Lipschitz of rank 1, has prescribed endpoints, and which maximizes the (unsigned) area enclosed between the graph of x and that of θ. The L^1 approximation problem may be summarized as follows:

$$\max \int_0^1 |x(t) - \theta(t)|\, dt \; : \; |x'(t)| \leqslant 1,\; x(0) = x_0,\; x(1) = x_1.$$

Prove that a solution of this problem exists. Identify the (unique) solution in the case $x_0 = x_1 = 0$. □

26.7 Exercise. The following control system, in which $n = 3$ and $m = 2$, is known as the *nonholonomic integrator*:

$$\begin{aligned} {x_1}' &= u_1 \\ {x_2}' &= u_2 \qquad\qquad (u_1, u_2) \in U := B(0,1) \subset \mathbb{R}^2 \\ {x_3}' &= x_1 u_2 - x_2 u_1 . \end{aligned}$$

It is the simplest representative of a class of systems arising in mechanics which have certain interesting properties.

(a) Let Π denote the (x_1, x_2) plane; that is, the surface $x_3 = 0$. Show that any initial value (α, β, γ) of the state for which $(\alpha, \beta) \neq 0$ can be steered to Π in finite time by a control $\hat{u}$ having the feedback form $\hat{u}(x_1, x_2, x_3) = \kappa(x_2, -x_1)$, for an appropriate value of the constant κ.

(b) Show that any point in Π can be steered to the origin in finite time, in such a way that the state remains in Π, by a constant control.

(c) Prove that the minimal-time problem for this system admits a solution.

(d) Prove that any time-optimal control is C^∞.

Remark. It can be shown that, despite these facts, there is no *continuous* feedback $(u_1, u_2) = \hat{u}(x_1, x_2, x_3) \in U$ which has the property that, from any initial condition, the state trajectories generated by $\hat{u}$ go to 0, even if we completely renounce time optimality. This is one reason why the system is often discussed in control engineering, in connection with feedback stabilization. □

26.8 Exercise. In the classical theory of economic production, there exist cases in which two inputs (for example, labor and capital) require *fixed proportions* in their use (one worker, one shovel). A simple model of this type is the case in which we dispose of two inputs u and v, which may be used to produce more capital x and more output y as follows:

$$x'(t) = u(t), \;\; y'(t) = \min(x(t), v(t)).$$

Thus, x and v must be used in equal proportions in producing the output. We shall interpret u and v as control variables whose values lie in the set

$$U = \{ (u,v) : u \geqslant 0, \; v \geqslant 0, \; u + v \leqslant 1 \}.$$

The initial values at time zero of the states x and y are taken to be zero, and the goal is to maximize total output

$$y(T) = \int_0^T \min\big(x(t), v(t)\big)\, dt,$$

where T is a prescribed horizon in $(0,1)$. Solve this optimal control problem. □

26.9 Exercise. We consider the following optimal control problem on a given interval $[0,T]$, a variant of the linear-quadratic regulator of Example 22.19:

$$\begin{cases} \text{Minimize} & J(x,u) = \int_0^T \Lambda\big(u(t)\big)\,dt \\ \text{subject to} & x'(t) = Ax(t) + Bu(t),\ t \in [0,T] \text{ a.e.} \\ & u(t) \in U := \mathbb{R}^m \\ & x(0) = x_0,\ x(T) = x_T \end{cases}$$

where $\Lambda(u) = \max\{0, |u|^2 - 1\}$. (Thus, control values $u \in B(0,1)$ are "free".) As before, we suppose that the controllability matrix $\mathscr{C}$ is of maximal rank.

(a) Prove that there is an optimal process (x_*, u_*).

Our goal is to characterize optimal processes. We observe that any control with values in the unit ball whose corresponding state trajectory joins x_0 to x_T in time T provides zero cost; it is evident that any such control is optimal.

(b) Show that by taking $|x_T - x_0|$ sufficiently large, we can be sure that no control with values exclusively in the unit ball can be admissible. We assume from now on that we are in this case.

(c) Show that the maximum principle applies, and that its necessary conditions must hold in normal form. Go on to show that the costate p is such that the zeros of the function $t \mapsto B^*p(t)$ (if any) are isolated.

(d) Prove that for almost every t we have

$$u_*(t) = \begin{cases} B^*p(t)/|B^*p(t)| & \text{if } 0 < |B^*p(t)| \leqslant 2 \\ B^*p(t)/2 & \text{if } |B^*p(t)| > 2\,. \end{cases}$$

(e) Deduce that u_* is continuous. □

26.10 Exercise. The Gordon-Schaefer *logistic model* in the theory of renewable resources postulates dynamics of the form

$$x'(t) = x(t)\big(\bar{x} - x(t)\big) - u(t)\,x(t)$$

where the state x measures the biomass of a given species and the control u corresponds to exploitation effort (for example, fishing). The net profit corresponding to an effort profile $u(t)$ is given by

$$\int_0^T e^{-\delta t}\{\,\pi x(t) - c\,\}\,u(t)\,dt\,,$$

where δ is the discount rate used to calculate the value of the revenue stream at time $t = 0$, π is the unit price of the resource, and c is the unit effort cost. The effort u

is restricted to the interval $[0,E]$. The carrying capacity $\bar{x}$ and the horizon T, along with δ, π, c, and E, are all given positive constants. The initial condition $x_0 \in (0,\bar{x})$ is prescribed. The resulting optimal control problem may be summarized as follows:

$$\begin{cases} \text{Maximize} & J(x,u) = \int_0^T e^{-\delta t}\{\pi x(t) - c\}\, u(t)\, dt \\ \text{subject to} & x'(t) = x(t)\big(\bar{x} - x(t)\big) - u(t)x(t) \\ & u(t) \in U := [0,E],\ t \in [0,T] \text{ a.e.} \\ & x(0) = x_0. \end{cases}$$

Prove that a solution exists. Deduce from the maximum principle that the solution is of turnpike type. (Note: in order for the turnpike to manifest itself, the planning horizon T must be sufficiently large, as well as the available effort E.) □

Optimal pricing. In introducing a product in the marketplace, one faces a decision regarding the initial price to charge.[1] Should one start low, so as to stimulate sales and create a fad effect, hoping to raise prices later? Or should one sell high to those willing to pay a lot, and then gradually lower the price so that each consumer pays the most that he /she is willing or able to? Can it be ever be optimal to have a price profile that exhibits a jump discontinuity, or is this inevitably a sign of non optimality? The basic model discussed here has done much to clarify questions such as these, and illustrates the role of necessary conditions in the qualitative analysis of solutions, even when these are not explicitly calculated.

We denote by x the (cumulative) quantity sold to the present time of a certain commodity, and u its price. The dynamics and initial value

$$x'(t) = Q\big(x(t), u(t)\big),\ x(0) = x_0\,,$$

describe the evolution of x, where Q is the *demand function*, which of course depends upon price, which plays the role of control variable. The function $C(x,q)$, which measures the *production cost*, is also given; it depends on past production x as well as the quantity (currently) produced q. Both C and Q are taken to be twice continuously differentiable, with $Q_u < 0$ and $Q > 0$. (These conditions have an evident interpretation.) The problem is to choose the price profile $u(t)$ so as to maximize the present value of the profit stream over a given interval $[0,T]$:

$$\int_0^T e^{-\delta t}\big\{u(t)\,Q\big(x(t),u(t)\big) - C\big(x(t), Q(x(t),u(t))\big)\big\}\, dt\,,$$

where u assumes values in $(0,\infty)$. The discount rate $\delta > 0$ is given.

---

[1] This exercise, and the next few, are based upon the results in the article *Optimal pricing policy in the presence of experience effects*, F. Clarke, M. Darrough, and J. Heineke, Journal of Business 55 (1982) 517-530.

26.11 Exercise. (Optimal pricing 1) Our first goal is to prove

Theorem. *Let u be an optimal price profile. Then, along the optimal process, we have*

$$u + Q/Q_u \;=\; C_q(x,Q) + \int_t^T \big(C_x(x,Q) + Q\,Q_x/Q_u\big)\, e^{-\delta(\tau-t)}\, d\tau\,.$$

(a) Why does the maximum principle apply in normal form? Show that the Hamiltonian H for the problem is given by

$$H(x,p,u) \;=\; \big(p + u\,e^{-\delta t}\big)\,Q(x,u) - e^{-\delta t}\,C\big(x,Q(x,u)\big).$$

(b) Prove that the adjoint variable p satisfies

$$p \;=\; e^{-\delta t}\big\{C_q(x,Q) - u - Q/Q_u\big\},\;\; t \in [0,T] \text{ a.e.}$$

(c) Prove that along the optimal process we have

$$C_{qq}(x,Q)\,Q_u^2 + Q\,Q_{uu}/Q_u - 2\,Q_u \;\geqslant\; 0.$$

(d) We define a function of (x,u) by

$$\varphi(x,u) \;=\; C_q\big(x,Q(x,u)\big) - u - Q(x,u)/Q_u(x,u).$$

and we set $\varphi(t) := \varphi(x(t),u(t))$. Prove that along the optimal process, we have $\varphi(t) = e^{\delta t}p(t)$ and

$$\varphi'(t) \;=\; \delta\varphi + Q\,Q_x/Q_u + C_x(x,Q),\;\; \varphi(T) = 0. \qquad (1)$$

(e) Deduce the theorem from these facts. Show also (for future reference) that we have:

$$\varphi_u \leqslant 0 \;\text{ along the optimal process.} \qquad (2)$$

The theorem can be interpreted in economic terms as affirming that the optimal price profile is such that long-run marginal cost and marginal revenue coincide. □

26.12 Exercise. (Optimal Pricing 2) The question of whether optimal price profiles in the problem of Exer. 26.11 are necessarily continuous is an important one. The following gives a sufficient condition for this.

Proposition. *Suppose that Q and C satisfy globally the inequality*

$$C_{qq}(x,Q)\,Q_u^2 + Q\,Q_{uu}/Q_u - 2\,Q_u \;>\; 0.$$

Then the optimal price path u is continuous and differentiable.

(a) Prove the proposition.

(b) Show that the hypothesis of the proposition is satisfied in the following circumstances: C is convex in q, and

$$Q(x,u) = f(x)\,g(u)\,, \text{ with } f > 0,\ g \text{ concave},\ g \geqslant 0,\ g' < 0.$$

(c) Show that when the optimal price profile u is differentiable, we have

$$u'\varphi_u = \delta\varphi + Q\,Q_x/Q_u + C_x(x,Q) - u\,\varphi_x \tag{3}$$

along the optimal process. □

26.13 Exercise. (Optimal Pricing 3) We now show in several natural cases how the theorem can be used to deduce the monotonicity properties of the optimal price profile. The first two of these cases concern what is called *learning by doing*, in which production costs decrease over time, because of the benefit of experience.

However, the exact nature of how the reduction takes place (by scaling or by translation) has a dramatic effect on the price path: in one case it decreases, in the other it increases.

We assume that an optimal price profile u exists, with u differentiable. (Results to come will identify criteria that imply this.)

(a) *Learning by doing*: *scaling in* C. We take C to be of the form

$$C(x,q) = c_0 + m(x)\,h(q)$$

where the functions m and h satisfy

$$m(x) > 0,\ m'(x) < 0,\ h \text{ is convex},\ h(0) = 0,\ h(q) > 0\ (q > 0),\ h'(q) > 0.$$

Thus the production cost curve is *scaled* downward as x increases. We also assume that demand is unaffected by experience: $Q_x = 0$. Prove with the help of (2) and (3) that in this scenario, an optimal price profile u is strictly decreasing.

(b) *Learning by doing*: *translation in* C. We now assume that C has the form

$$C(x,q) = c_0 + s(x) + r(q),$$

where we have $r' > 0,\ s' < 0$, and s is convex on $(0,\infty)$. Thus, experience now *translates* the cost curve downwards. We continue to assume $Q_x = 0$. Prove that in this scenario, an optimal price profile u is strictly increasing.

(c) *Demand experience.* Let $C(x,q) = c_0\,q$ and let the demand function be of the form $Q(x,u) = \sigma(x)\,\rho(u)$, where σ and ρ are positive and bounded away from 0, with $\rho' < 0$. We further assume that for a certain $\bar{x} > x_0$, we have

$$\sigma'(x) = 0 \text{ for } 0 < x < \bar{x}, \quad \sigma'(x) < 0 \text{ for } x > \bar{x}.$$

(Thus, $\bar{x}$ is a threshold beyond which consumer satiation sets in.) Note that for T sufficiently large, the state x breaches this threshold. Prove then that the optimal price path exhibits at least one period of increase, and at least one period of decrease. □

26.14 Exercise. (Optimal Pricing 4) With reference to the optimal pricing problem of Exer. 26.11, we shall construct an explicit example in which the optimal price path is discontinuous (a conclusion that economists find surprising). We take

$$x_0 = 0, \; T = 1, \; Q(x,u) = e^{-u/100}.$$

Note that x and Q lie in $[0,1]$ at all times. Let $\alpha < \beta < \gamma$ be three positive numbers, and let f be any continuously differentiable function satisfying

(1) $f(q) \geqslant 0$ for $q \in [0,1]$; $f(q) = 0$ if and only if $q = \beta$ or $q = \gamma$.

(2) $|f'(q)| \leqslant 1$ for $q \in [0,1/3]$.

(3) $f'(q) \geqslant 1$ for $q \in [1/3, 3/4]$.

(4) $f'(q) > 104$ for $q \in [3/4,1]$.

We introduce the cost function $C(x,q) = -100\, q \ln q + f(q) + g(x,q)$, where

$$g(x,q) = \big[\max \big\{ (x-\alpha)(q-\beta)^2, (\alpha-x)(q-\gamma)^2 \big\} \big]^2.$$

(a) Show that g is continuously differentiable, as is $q \ln q$ for $q > 0$.

(b) Prove that, under conditions (1) to (4), we have $C_q > 0$. (Thus, on the face of it, C is not unreasonable as a cost function.)

(c) Show that the optimal pricing problem becomes that of minimizing

$$\int_0^1 e^{-\delta t} h\big(x(t), x'(t)\big)\, dt, \text{ where } h := f + g.$$

(d) Observe that h is nonnegative by construction, and can only equal zero in two ways: $q = \gamma$ and $x \leqslant \alpha$, or $q = \beta$ and $x \geqslant \alpha$. Deduce that the unique optimal policy is given by

$$x' = e^{-u/100} = \begin{cases} \gamma & \text{until } x = \alpha \\ \beta & \text{thereafter.} \end{cases}$$

Thus the optimal price profile is piecewise constant, with an upward jump at time $t = \alpha/\gamma$. □

26.15 Exercise. (Optimal Pricing 5) One may have a guilty conscience about simply assuming the existence of an optimal price profile in Exer. 26.11. (In fact, we hope the reader has learned to feel uneasy when this question is simply ignored.) The fact that the control set $U = (0,\infty)$ is open makes *a priori* existence hard to prove in general.

With the help of the necessary conditions provided by the extended maximum principle 22.26, we shall establish an existence theorem that applies when the demand function has the form $Q(x,u) = \alpha e^{-\beta u}$, where α, β are positive constants and $\alpha < 1$. In this setting, the goal is to prove:

Theorem. *Suppose that for some constant* $m \geqslant 0$*, the cost function* C *satisfies*

$$C_{qq} \geqslant -1/(\alpha\beta)\,, \;\; C_x \geqslant -m\,, \;\; C_q \geqslant m e^{\delta T}/\delta - 1/\beta\,.$$

Then there is an optimal path profile u *having values in* $(0,\infty)$*, and* u *is continuous and differentiable.*

(a) Prove that the problem may be recast as that of minimizing

$$\int_0^T e^{-\delta t}\big\{C\big(x(t),x'(t)\big)+x'(t)\ln\big(x'(t)/\alpha\big)/\beta\,\big\}\,dt$$

subject to the initial condition on x, and the constraint $0 < x'(t) < \alpha$.

(b) Prove that the Lagrangian $\Lambda(t,x,v)$ of this problem is convex in v.

(c) Prove by the direct method that the problem above admits a solution if the inequality restricting x' is extended to $x'(t) \in [0,\alpha]$.

If the solution x of the extended problem is such that $0 < x'(t) < \alpha$ a.e., then it is clear that x is also a solution of the original problem. We proceed to establish this.

(d) Show that the necessary conditions of Theorem 22.26 apply to the extended problem, and imply the existence of an arc p with $p(T) = 0$ such that, for almost every t in $[0,T]$:

$$p'(t) = e^{-\delta t}C_x\,, \quad \min_{0 \leqslant v \leqslant \alpha}\big\{\Lambda\big(t,x(t),v\big) - p(t)\,v\big\} \text{ is attained at } v = x'(t).$$

(e) Deduce from this that $x'(t) > 0$ a.e.

(f) Use the hypotheses on C to show that $p(t) < m/\delta$ for every $t \in [0,T]$.

(g) Now suppose that $x'(t) = \alpha$ on a set of positive measure. Deduce from (d) above that, for some $\tau \in (0,T)$, we have $p(t) \geqslant m/\delta$, which contradicts (f). It follows that x solves the original problem.

(h) Finally, prove that the optimal path profile u is continuous and differentiable. □

26.16 Exercise. The goal is to prove Theorem 25.17. Let (x,u) be any admissible process for the problem (MC).

(a) Show that $J(x,u)-J(x_*,u_*)$ is bounded below by

$$\nabla\ell\big(x_*(a),x_*(b)\big)\bullet\big(x(a)-x_*(a),x(b)-x_*(b)\big) \\ +\int_a^b \Lambda_{x,u}\big(t,x_*,u_*\big)\bullet(x-x_*,u-u_*)\,dt\,,$$

and go on to show that the first term above is bounded below by

$$\big(p(a),-p(b)\big)\bullet\big(x(a)-x_*(a),x(b)-x_*(b)\big).$$

(b) Substitute for Λ_x and Λ_u from the adjoint equation and the stationarity condition, in order to bound from below the integral term in the expression above.

(c) Derive from these lower bounds that $J(x,u)-J(x_*,u_*)$ is nonnegative, which yields the theorem. □

26.17 Exercise. Let $U(\cdot)$ be a multifunction from $[a,b]$ to the subsets of $\mathbb{R}^m$, and define C to be the set of all measurable functions u mapping $[a,b]$ to $\mathbb{R}^m$ such that $u(t)\in U(t)$ a.e. It can be proved[2] that C is a complete metric space when equipped with the following metric d:

$$d(u,v)\;=\;\text{meas}\big\{t\in[a,b]\,:\,u(t)\neq v(t)\big\}.$$

Let the dynamics function f and running cost Λ be LB measurable in t and (x,u), and have linear growth: for a summable function M, for almost every t, we have

$$|f(t,x,u)|+|\Lambda(t,x,u)|\leqslant M(t)\big(1+|x|\big)\;\;\forall\,t\in[a,b],\;x\in\mathbb{R}^n,\;u\in U(t)\,.$$

We also assume that f is Lipschitz in x, as follows:

$$t\in[a,b],\,u\in U(t)\implies|f(t,x_1,u)-f(t,x_2,u)|\leqslant k|x_1-x_2|\;\;\forall\,x_1,x_2\in\mathbb{R}^n.$$

Let $\ell:\mathbb{R}^n\to\mathbb{R}$ be lower semicontinuous, and fix $x_0\in\mathbb{R}^n$. Then every control $u\in C$ generates a unique state trajectory x_u of the control system (f,U) on $[a,b]$ with initial condition $x(a)=x_0$. (Why?) Prove that the cost functional

$$J(u)\;=\;\ell\big(x_u(b)\big)+\int_a^b\Lambda\big(t,x_u(t),u(t)\big)\,dt$$

is well defined and lower semicontinuous on the metric space C. □

[2] See Clarke [13, p. 202].

26.18 Exercise. The purpose of this exercise is to give a precise meaning to the statement that the set of relaxed trajectories (see § 23.1) is the closure of the set of original trajectories. We prove the following approximation result:

Theorem. *Let (f,U) be a control system on the interval $[a,b]$ such that $f(t,x,u)$ is continuous in (x,u) and measurable in t, and $U(\cdot)$ is measurable and compact valued. Let y be a relaxed trajectory of the system (f,U) on the interval $[a,b]$. Suppose that, for some $\delta > 0$ and constant k, for almost every t, we have the following Lipschitz condition*:

$$|x_i - y(t)| \leqslant \delta \ \ (i = 1,2),\ u \in U(t) \implies |f(t,x_1,u) - f(t,x_2,u)| \leqslant k|x_1 - x_2|.$$

Then, for any $\varepsilon > 0$, there exists an original trajectory x of the system such that

$$x(a) = y(a),\ |x(t) - y(t)| \leqslant \varepsilon \ \forall t \in [a,b].$$

We remark that the Lipschitz property is an essential hypothesis for the result. The usual proofs, which are somewhat involved, invoke the Lyapunov convexity theorem or one of its variants. We give here a guided proof based upon the extended maximum principle.

A. We begin with a useful extension device which allows us to prove the theorem under hypotheses of a global, rather than local, nature. Let $\pi_t : \mathbb{R}^n \to \mathbb{R}^n$ be the projection onto the set $B(y(t),\delta)$, and, for any x, set $f(t,x,u) = f(t,\pi_t x,u)$. As concerns the x variable, this uses only the values of f on the set $B(y(t),\delta)$, and leaves f unchanged there. Show that the new function f defined in this way satisfies the following *global* linear growth and Lipschitz conditions: for some summable function M, for almost every $t \in [a,b]$,

$$x \in \mathbb{R}^n,\ u \in U(t) \implies |f(t,x,u)| \leqslant M(t)(1+|x|),$$
$$x_1,x_2 \in \mathbb{R}^n,\ u \in U(t) \implies |f(t,x_1,u) - f(t,x_2,u)| \leqslant k|x_1 - x_2|.$$

Show that the validity of the theorem for the redefined system implies its validity for the original one. We therefore assume that the global conditions above hold.

B. Show that it suffices to prove the following variant of the theorem (in which the hypotheses are unchanged): for any $\varepsilon > 0$, there exists an original trajectory x of the system (f,U) such that

$$x(a) = y(a),\quad \int_a^b |x(t) - y(t)|\,dt \leqslant \varepsilon.$$

C. Show that, in order to prove the variant, it suffices to establish the existence of a number $\rho > 0$ with the following property: let $[a_1,b_1]$ be any subinterval of $[a,b]$ having $b_1 - a_1 \leqslant \rho$; then, given any $\varepsilon > 0$ and $\alpha \in \mathbb{R}^n$ sufficiently small, there exists a trajectory x for (f,U) on $[a_1,b_1]$ that satisfies

$$x(a_1) = y(a_1)+\alpha, \quad \int_{a_1}^{b_1} |x(t)-y(t)|\,dt \leqslant \varepsilon.$$

The remaining steps will prove the existence of such a ρ.

D. For a fixed subinterval $[a_1,b_1]$ and $\alpha \in \mathbb{R}^n$, consider the optimal control problem:

$$\inf \int_{a_1}^{b_1} |x(t)-y(t)|\,dt,\ x(a_1) = y(a_1)+\alpha,$$

where the infimum is taken over all the trajectories x of (f,U) on $[a_1,b_1]$. The hypotheses do not imply that the infimum is attained. Show, however, with the help of Exer. 26.17, that for any $\varepsilon > 0$, there is a process $(x_\varepsilon,u_\varepsilon)$ that minimizes over all processes (x,u) for (f,U) on $[a_1,b_1]$ the cost

$$\int_{a_1}^{b_1} \{\,|x(t)-y(t)|+\varepsilon\,\theta(u(t)-u_\varepsilon(t))\}\,dt,$$

where θ is the function on $\mathbb{R}^m$ which equals 0 at 0 and 1 everywhere else.

E. Apply the extended maximum principle to deduce the existence of an arc p on $[a_1,b_1]$ satisfying $p(b_1) = 0$ and, for almost every t:

$$|p'(t)| \leqslant k|p(t)|+1,$$
$$\langle p'(t), y(t)-x_\varepsilon(t)\rangle \leqslant k|p(t)|\,|y(t)-x_\varepsilon(t)|-|x_\varepsilon(t)-y(t)|,$$
$$\langle p(t), f(t,x_\varepsilon(t),u)\rangle \leqslant \langle p(t), x_\varepsilon'(t)\rangle+\varepsilon\ \ \forall u \in U(t).$$

F. Deduce first that

$$\langle p(t), y'(t)\rangle \leqslant \langle p(t), x_\varepsilon'(t)\rangle + k|p(t)|\,|y(t)-x_\varepsilon(t)|+\varepsilon \text{ a.e.},$$

and then prove

$$\frac{d}{dt}\langle p(t), y(t)-x_\varepsilon(t)\rangle \leqslant 2k|p(t)|\,|y(t)-x_\varepsilon(t)|+\varepsilon-|x_\varepsilon(t)-y(t)| \text{ a.e.}$$

Suppose now that $\rho > 0$ is chosen small enough so that

$$b_1-a_1 \leqslant \rho,\ p(b_1) = 0,\ |p'(t)| \leqslant k|p(t)|+1 \text{ a.e.}$$
$$\Longrightarrow\ 2k|p(t)| \leqslant 1/2\ \ \forall t \in [a_1,b_1].$$

Proceed to discover

$$(1/2)\int_{a_1}^{b_1} |x_\varepsilon(t)-y(t)|\,dt \leqslant \varepsilon\rho+|\alpha|/(4k),$$

and conclude. ☐

Control Lyapunov functions. The next set of four exercises deals with an autonomous control system (f,U) satisfying linear growth: for certain constants c and d, we have

$$|f(x,u)| \leqslant c|x|+d \;\; \forall x \in \mathbb{R}^n, \;\; \forall u \in U.$$

We suppose in addition that f is locally Lipschitz, U is compact, and the velocity set $f(x,U)$ is convex for each x. We say that the system is *null controllable* if, for all $\alpha \in \mathbb{R}^n$, there exists a state trajectory x defined on a finite interval $[0,T]$ such that $x(0)=\alpha$ and $x(T)=0$. We define a *control Lyapunov function* to be a lower semicontinuous function $\varphi : \mathbb{R}^n \to \mathbb{R}_+$ satisfying the following properties:

Positive definiteness: $\varphi(0)=0$ and $\varphi(x)>0 \;\; \forall x \neq 0$.

Infinitesimal decrease: For some $\omega>0$, we have

$$\min_{u \in U} \, d\varphi\big(x; f(x,u)\big) < -\omega \;\; \forall x \neq 0.$$

Here, as usual, $d\varphi$ refers to the Dini derivate (see Def. 11.18); we are limiting attention to a *constant* decrease rate ω (independent of x) for simplicity. When φ is smooth, infinitesimal decrease may be written in the equivalent form

$$\min_{u \in U} \, \langle \nabla\varphi(x), f(x,u)\rangle < -\omega \;\; \forall x \neq 0.$$

We have seen earlier the role of Lyapunov functions in guaranteeing the stability of an *uncontrolled* differential equation (see Prop. 12.18). In the setting of a control system, it is controllability that we seek to confirm.

26.19 Exercise. (Lyapunov 1) Under the hypotheses above, the goal is to prove:

Theorem. *If the system admits a control Lyapunov function in the above sense, then it is null controllable.*

Let $\alpha \neq 0$. We wish to establish the existence of a trajectory x on an interval $[0,T]$ such that $x(0)=\alpha$ and $x(T)=0$. Prove the existence of a trajectory x on an interval $[0,\infty)$ such that

$$\varphi\big(x(t)\big)+\omega t \leqslant \varphi(\alpha)$$

as long as $t>0$ is such that $x(s) \neq 0$ for $0 \leqslant s \leqslant t$. Deduce the theorem from this, as well as an estimate on the minimal time $T(\alpha)$ needed to reach 0. □

26.20 Exercise. (Lyapunov 2) Prove that (under the same hypotheses as above) if the system is null controllable, then the minimal-time function $T(\cdot)$ is a control Lyapunov function.[3] (Here, of course, $T(\alpha)$ is the least time taken by a trajectory to join α to 0.) □

[3] This gives meaning to the statement: *the control system is null controllable if and only if there exists a control Lyapunov function.* Note, however, that *nonsmooth* Lyapunov functions are used to formulate this principle; it is false otherwise.

For purposes of stabilization and feedback, topics that we do not address here, it is desirable to have a control Lyapunov function φ that is continuous.

26.21 Exercise. (Lyapunov 3) Suppose that the system admits a *continuous* control Lyapunov function. Prove that the minimal-time function $T(\cdot)$ is continuous at 0, and that $0 \in f(0,U)$. □

26.22 Exercise. (Lyapunov 4) Assume that $0 \in f(0,U)$. Prove:

Theorem. *The following are equivalent*:

(a) *There exists a continuous control Lyapunov function* φ;

(b) *The minimal-time function* $T(\cdot)$ *is continuous at* 0;

(c) *The minimal-time function* $T(\cdot)$ *is continuous.* □

Controllability and normality. Let the system (f,U) satisfy the classical smoothness hypotheses. A process (x_*,u_*) of the system on the interval $[a,b]$ is called *normal* if the only costate p satisfying the adjoint equation and the maximum condition of Theorem 22.2 abnormally (that is, with $\eta = 0$) is the zero arc. Note that this property is unrelated to any kind of optimality.

The remaining exercises in this chapter share the following:

Standing hypotheses: (f,U) is a finitely generated autonomous system (see Def. 23.8), where the vector fields g_0 and g_j are continuously differentiable and have linear growth, and where U is compact and convex.

26.23 Exercise. (Controllability 1) The goal is to prove that the system is locally controllable around a normal process, in the following sense:

Theorem. *Let* (x_*,u_*) *be a normal process on the interval* $[a,b]$. *Then there exist constants* K *and* $\delta > 0$ *with the following property*: *for every* $\alpha,\beta \in B(0,\delta)$, *there exists a process* (x,u) *of the system satisfying*

$$x(a) = x_*(a)+\alpha\,,\ \ x(b) = x_*(b)+\beta \tag{4}$$

as well as

$$\int_a^b \{\,|x(t)-x_*(t)|+|u(t)-u_*(t)|\,\}\,dt \ \leqslant\ K|(\alpha,\beta)|.$$

(a) We define the value function

$$V(\alpha,\beta) = \min \int_a^b \{\,|x(t)-x_*(t)|+|u(t)-u_*(t)|\,\}\,dt\,,$$

where the minimum is taken over the processes (x,u) of the system satisfying (4). The strategy of the proof is to show that V is Lipschitz near $(0,0)$. Show that this property gives the required conclusion.

(b) Prove that the minimum defining $V(\alpha,\beta)$ is attained when $V(\alpha,\beta) < \infty$; that is, when there exists at least one admissible process corresponding to (α,β).

(c) Prove that V is lower semicontinuous.

(d) Let $(\zeta,\psi) \in \partial_P V(\alpha,\beta)$, and let (x,u) be a process attaining the minimum in the definition of $V(\alpha,\beta)$. Show that there exists an arc p and $\eta \in \{0,1\}$ such that

$$p(a) = -\zeta\,, \quad p(b) = \psi\,, \quad -p'(t) \in D_x f^*\big(x(t),u(t)\big) + \eta\, B \text{ a.e.},$$

$$\max_{w \in U} \langle p(t), f(x(t),w)\rangle - \eta\,|w - u_*(t)| \text{ at } w = u(t) \text{ a.e.}$$

(e) Suppose that there is a sequence $(\zeta_i,\psi_i) \in \partial_P V(\alpha_i,\beta_i)$ which is unbounded, where $(\alpha_i,\beta_i) \to (0,0)$ and $V(\alpha_i,\beta_i) \to V(0,0) = 0$. Prove that this leads to the conclusion that the process (x_*,u_*) is *not* normal (a contradiction).

(f) Deduce that V Lipschitz near $(0,0)$. □

We suppose in the next four exercises that (in addition to the standing hypotheses) we have $0 \in U$ and $f(0,0) = 0$. Then the zero control corresponds to the state trajectory $x \equiv 0$; thus, $(0,0)$ is a process of the system on $[0,T]$, for any $T > 0$. We say that *the origin is normal* if there exist horizons T arbitrarily small such that the process $(0,0)$ is normal on $[0,T]$. As we now see, this property implies null controllability.

26.24 Exercise. (Controllability 2) Prove that if the origin is normal, then the minimal-time function is continuous. □

26.25 Exercise. (Controllability 3) Prove that if $0 \in \operatorname{int} f(0,U)$, then the origin is normal. □

26.26 Exercise. (Controllability 4) Let G and g_0 be given as in Def. 23.8, and set $A = Dg_0(0)$, $B = DG(0)$. Suppose that

$$\varepsilon > 0\,,\;\; q \in \mathbb{R}^n,\;\; B^* e^{-A^* t} q \;\in N_U(0)\;\; \forall t \in [0,\varepsilon] \;\implies\; q = 0.$$

Show that the origin is normal. □

26.27 Exercise. (Controllability 5) Let A and B be as in the preceding exercise, and let $\mathscr{C}$ be the controllability matrix defined as in Example 22.19. Show that if $0 \in \operatorname{int} U$ and $\mathscr{C}$ has maximal rank, then the origin is normal. We remark that an important special case of this occurs in classical linear systems theory, in which

$$f(x,u) = Ax + Bu\,,\;\; 0 \in \operatorname{int} U\,,\;\; \mathscr{C} \text{ has maximal rank.}$$

□

Sensitivity. The next two exercises study the sensitivity of control systems with respect to the initial condition. The standing hypotheses remain in force: (f,U) is an autonomous finitely generated system, where the vector fields g_0 and g_j are continuously differentiable and have linear growth, and where U is compact and convex.

26.28 Exercise. (Sensitivity 1) Let S be a compact subset of $\mathbb{R}^n$ and let $T_0 > 0$. The goal is to prove the following.

Theorem. *There exists K such that, for any $T \in (0,T_0]$, for any trajectory x of the system on the interval $[0,T]$ having $x(0) \in S$, for any $\alpha \in B(0,1)$, there is a trajectory y on $[0,T]$ satisfying $y(0) = x(0) + \alpha$ and*

$$|x(T) - y(T)| \leqslant K|\alpha|.$$

(a) Let Σ be any bounded subset of $\mathbb{R}^n \times [0,\infty)$. Prove that there exists $M(\Sigma)$ such that, for any trajectory x on an interval $[0,T]$ with $(x(0),T) \in \Sigma$, we have $|x(t)| \leqslant M$. (Use Gronwall's lemma.)

(b) Fix x and T as in the statement of the theorem. Consider on $B(0,1)$ the function

$$V(\alpha) = \min |x(T) - y(T)|,$$

where the minimum is taken over all trajectories y satisfying $y(0) = x(0) + \alpha$. Show that V is finite and lower semicontinuous, and that the minimum defining $V(\alpha)$ is attained when $V(\alpha) < \infty$.

(c) Let $\zeta \in \partial_P V(\alpha)$, and let y_α solve the problem defining $V(\alpha)$. Interpret the proximal subgradient inequality to discover that y_α solves a certain optimal control problem in which the cost contains a term of the form $\langle -\zeta, y(0)\rangle$, and in which $y(0)$ is unconstrained.

(d) Apply the extended maximum principle to the problem in question, and obtain an *a priori* bound on ζ from the necessary conditions.

(e) Deduce from this that V is Lipschitz on $B(0,1)$, and conclude. □

26.29 Exercise. (Sensitivity 2) The reader is familiar with the interpretation of the multiplier that appears in the multiplier rule as a measure of sensitivity. In this exercise, we see how the costate p of the maximum principle may be viewed in this light. Consider the following optimal control problem on $[0,1]$:

$$\left\{\begin{array}{ll} \text{Minimize} & J(x,u) = \ell\big(x(1)\big) + \int_0^1 \Lambda\big(u(t)\big)\,dt \\ \text{subject to} & x'(t) = f\big(x(t),u(t)\big) \text{ a.e.} \\ & u(t) \in U \text{ a.e.} \\ & x(0) = x_0 + \alpha\,. \end{array}\right. \qquad (\mathrm{P}_\alpha)$$

Here, the point x_0 and the continuously differentiable functions $\ell : \mathbb{R}^n \to \mathbb{R}$ and $\Lambda : \mathbb{R}^m \to \mathbb{R}$ are given; we further assume that Λ is convex. The role of $\alpha \in \mathbb{R}^n$ is that of a parameter. Our interest centers around the function

$$V(\alpha) = \min\,(\,\mathrm{P}_\alpha\,)$$

for α near 0; this measures the effect of perturbing the initial condition. We denote by $\Sigma(\alpha)$ the set of optimal processes (x,u) for the problem $(\,\mathrm{P}_\alpha)$.

(a) Prove that $\Sigma(\alpha)$ is nonempty for each α, and that, for any $(x,u) \in \Sigma(\alpha)$, there exists a unique costate p satisfying $p(1) = -\nabla\ell(x(1))$ and

$$-p'(t) = D_x H\big(x(t), p(t), u(t)\big) \text{ a.e.}, \quad H\big(x(t), p(t), u(t)\big) = M\big(x(t), p(t)\big) \text{ a.e.}$$

We denote by $p_{x,u}$ the costate p that corresponds as above to the process (x,u). Here is the result we are aiming for:

Theorem. *The function V is locally Lipschitz, and we have*

$$\big\{ -p_{x,u}(0) : (x,u) \in \Sigma(0)\big\} \supset \partial_L V(0) \neq \emptyset.$$

The first two steps in the proof outlined below follow in essentially the same way as in Exer. 26.28.

(b) Let $(x,u) \in \Sigma(\alpha)$, and let $\zeta \in \partial_P V(\alpha)$. Prove that $p_{x,u}(0) = -\zeta$.

(c) Prove that V is locally Lipschitz.

(d) Prove the theorem.

(e) Show that if $(\,\mathrm{P}_0\,)$ admits a unique solution (x,u), then V is differentiable at 0; find $\nabla V(0)$.

(f) Identify additional hypotheses on the data that guarantee the preceding case. □

26.30 Exercise. (Hamilton-Jacobi-Bellman equation) The goal of this exercise is to characterize the (unique generalized) solution of a form of the Hamilton-Jacobi equation that arises in optimal control theory. This is in the same line of thought, then, as the results in § 19.3 (p. 379) and § 24.3 (p. 500). Let the system (f,U) satisfy the current standing hypotheses, and let $\ell : \mathbb{R}^n \to \mathbb{R}$ be continuous.

Our interest centers upon the solutions φ of the following partial differential equation with boundary condition:

$$\varphi_t(t,x) + h\big(x, \varphi_x(t,x)\big) = 0 \;\; \forall\,(t,x) \in \Omega, \quad \varphi(T,x) = \ell(x) \;\; \forall\, x \in \mathbb{R}^n, \tag{5}$$

where $\Omega = (-\infty, T) \times \mathbb{R}^n$, where h is the lower Hamiltonian of the system:

$$h(x,p) = \min_{u \in U} \langle\, p, f(x,u)\rangle,$$

and where ℓ is a continuous function. Since there fails to be a classical (smooth) solution φ of (5) in general, we require an extension of the solution concept which allows nonsmooth functions φ.

Accordingly, we say that a continuous function $\varphi : (-\infty, T\,] \times \mathbb{R}^n \to \mathbb{R}$ is a *proximal solution* of (5) if it satisfies $\varphi(T, x) = \ell(x)\ \forall x \in \mathbb{R}^n$ as well as

$$\theta + h(x, \zeta) = 0 \ \forall (\theta, \zeta) \in \partial_P \varphi(t, x), \ \forall (t, x) \in \Omega . \tag{6}$$

The goal is to prove the following:

Theorem. *There exists a unique proximal solution of* (5)*, namely the function*

$$V(\tau, \alpha) = \min \ \ell(x(T)),$$

where the minimum is taken with respect to all state trajectories x on $[\,\tau, T\,]$ that satisfy $x(\tau) = \alpha$.

(a) Prove that V is continuous. (Exer. 26.28 may be of some help.)

(b) Prove that V satisfies (6). [Hint: system monotonicity.]

(c) Show that V is the only proximal solution of (5). □

Notes, solutions, and hints

The trouble with a book is that you never know what's in it until it's too late.
Jeanette Winterson (Why Be Happy When You Could Be Normal?)

Concentration of energy, that was what he found in their books; a willingness to save someone else the time they had themselves expended.
Hugh Kenner (The Mechanic Muse)

This morning I took out a comma, and this afternoon I put it back in again.
Oscar Wilde

Part I. Functional analysis.

In the preface, the author mentioned his gratitude towards the teachers who contributed to his mathematical education. In these endnotes, we wish to acknowledge the beneficial influence of certain notable books. As regards functional analysis, we cite the celebrated texts of Royden [36] and Rudin [37, 38], which have been constant companions. Among our other nominees for books to bring to a desert island, there would certainly be those of Dunford and Schwartz [21] and of Edwards [22], which have quasi-biblical status in the field.

We also make a few suggestions for parallel or further reading. For convex analysis, we recommend Rockafellar [34] and Hiriart-Urruty and Lemaréchal [29]. Variational principles, and their links to norm smoothness and other properties of normed spaces, are studied in Deville, Godefroy and Zizler [20]; see also Phelps [32]. The elegant book of Brézis [8] develops functional analysis with an eye to applications in partial differential equations; see also Aubin [2] and Aubin and Ekeland [3], which stress instead equilibria and optimization.

We now list some partial solutions and hints for selected exercises in Part I.

1.35 Consider taking $L = \mathbb{R}x_0$ and $\lambda(tx_0) = t\|x_0\|^2$ in the context of Cor. 1.33.

1.42 (b): By Exer. 1.38, the normal cone is $\mathbb{R}_- \times \{0\} \times \mathbb{R}_+$.

2.8 By Carathéodory's theorem, $\text{co}(A)$ is the image of the compact set

$$\left\{ (t_0, t_1, \dots, t_n, x_0, x_1, \dots, x_n) : t_i \geqslant 0,\ \sum t_i = 1, x_i \in A \right\}$$

under the continuous function $f(t_0, t_1, \dots, t_n, x_0, x_1, \dots, x_n) = \sum_i t_i x_i$.

2.14 Without supposing that the infimum is finite, let us suppose that it fails to be attained. Let x_n be a sequence such that $f(x_n)$ decreases strictly to the infimum. Then $\{x : f(x) > f(x_n)\}$ is an open covering of E, so it admits a finite subcover. It follows that $\inf_E f$ is the last value $f(x_n)$ among the elements of the subcover, a contradiction.

2.28 The second fails to be convex. The determinant of $\nabla^2 f$, which is the product of its eigenvalues, is $-e^{2xy}(1+2xy)$. When this is negative, $\nabla^2 f$ has a negative eigenvalue.

F. Clarke, *Functional Analysis, Calculus of Variations and Optimal Control*, Graduate Texts in Mathematics 264, DOI 10.1007/978-1-4471-4820-3,

2.32 Each $x \in C$ admits $r(x) > 0$ and $K(x)$ such that f is Lipschitz of rank $K(x)$ on $B(x, 2r(x))$. Let $\{B(x_i, r(x_i)\}$ be a finite collection covering C, and consider

$$K > \max \left[\max_i K(x_i), 2M/(\min_i r(x_i)) \right],$$

where $|f| \leqslant M$ on C.

2.41 For (c) implies (a): observe that $(1,0,0,\dots,0)$ does not lie in the set

$$\{(f_0, f_1, \dots, f_n)(x) : x \in E\}.$$

This is a subspace of $\mathbb{R}^{n+1}$; thus, a closed convex set. Separate.

2.44 Show that for any $p \in \mathbb{R}^n$, we have $H_D(p) \geqslant \langle p, m \rangle$, where

$$m = \frac{1}{b-a} \int_a^b f(t)\,dt\,.$$

2.45 If $C+S \subset D+S$, then $H_C + H_S = H_{C+S} \leqslant H_{D+S} = H_D + H_S$. We may subtract H_S (because H_S is finite-valued) to deduce $H_C \leqslant H_D$, whence $C \subset D$ by Prop. 2.42.

2.49 For the last part, recall that any $\zeta \in N_C(0)$ can be identified with an element of ℓ^2.

3.8 The map $x \mapsto \|x\|$ is convex and strongly lsc (since continuous); then it is weakly lsc by Cor. 3.7; the result follows.

3.10 The point x, as the weak limit of the sequence x_i, lies in the weak closure of C. Then, by Theorem 3.6, it lies in the strong closure.

3.22 One may use Cor. 3.15, calling upon Theorem 3.21 to obtain sequential compactness.

4.9

$$\varphi\left(\int_\Omega g(x)\,dx\right) \leqslant \frac{1}{\text{meas}\,\Omega} \int_\Omega \varphi\big(g(x)\big)\,dx \quad \forall g \in L^1(\Omega)\,.$$

4.11 This follows from applying Theorem 4.10 to I_C and I_D (see Exer. 4.5). The intersection formula can be asserted under weaker hypotheses; see Exer. 8.23.

4.17
(a) Suppose that the strict subgradient condition holds. Let $x \neq y$ be given, fix $t \in (0,1)$, and set $z = (1-t)x + ty$. Let $\zeta \in \partial f(z)$ (which we know to be nonempty). Then $f(x) - f(z) > \langle \zeta, x - z \rangle$ and $f(y) - f(z) > \langle \zeta, y - z \rangle$. Dividing these inequalities by t and $1-t$ respectively and adding, we arrive at $f(z) < (1-t)f(x) + tf(y)$, which verifies the strict convexity.

Conversely, let f be strictly convex, and let $x \neq y$, $\zeta \in \partial f(x)$ be given. Set z equal to $(x+y)/2$. Then $f(z) < (f(x) + f(y))/2$, which leads to

$$f(y) - f(x) > 2\big(f(z) - f(x)\big) \geqslant 2\langle \zeta, z - x \rangle = \langle \zeta, y - x \rangle.$$

(b) That ∂f is injective when f is strictly convex follows easily by contradiction, using (a). If f is not strictly convex, then by (a) we have $\zeta \in \partial f(x)$ and $x \neq y$ such that $f(y) - f(z) = \langle \zeta, y - z \rangle$. Then x minimizes $f(u) - \zeta \bullet u$. But this function has the same value at y as at x, whence $\zeta \in \partial f(y)$, which shows that ∂f is not injective.

4.24 Hint for (b): the function $x \mapsto \langle \zeta, x \rangle - H_\Sigma(x)$ is positively homogeneous; Cor. 3.13 will also play a role.

4.38 We may assume that $\text{dom}\, g$ is nonempty. Then the assertion is a direct consequence of Theorem 4.36, with $U = kB_*$ (weak* topology) and $V = \text{dom}\, g$.

5.4 Here is a sketch of the completeness proof for ℓ^1. Let x^j be Cauchy in ℓ^1; then there exists M such that $\|x^j\|_1 \leqslant M \ \ \forall j$. Show that there is a sequence $x = (x_1, x_2, \dots)$ such that, for each i, $\lim_j x_i^j = x_i$. Prove that $\|x\|_1 \leqslant M$. There remains to show that $\|x^j - x\|_1 \to 0$. Let $\varepsilon > 0$. Choose k so that $\|x^j - x^k\|_1 < \varepsilon \ \ \forall j \geqslant k$, and m so that $\sum_{i \geqslant m} |x_i^k - x_i| < \varepsilon$. It follows that, for all j sufficiently large, we have $\|x^j - x\|_1 < 3\varepsilon$.

5.9 The exercise amounts to proving what is referred to as the du Bois-Raymond lemma; the gist of the argument appears in the proof of Theorem 15.2.

5.10 Exer. 5.8 is helpful here.

5.13 Let A be a weakly compact set. Each point x of A induces an operator in $L_C(X^*, \mathbb{R})$ by the formula $x(\zeta) = \langle \zeta, x \rangle$, one whose norm is $\|x\|_X$. Because A is weakly compact, the family of these operators is simply bounded on X^*, which is a Banach space. By the uniform boundedness principle, the set $\{\, \|x\|_X : x \in A \,\}$ is bounded.

5.16 For each positive integer i, define a map ζ_i on ℓ_c^∞ by $\langle \zeta_i, x \rangle = i x_i$.

5.17 For (a): by Cor. 2.35, the function $g(t) = f(x_0 + tz)$ is continuous at 0.

5.27 Consider the identity map $(X, \|\cdot\|_2) \mapsto (X, \|\cdot\|_1)$.

5.40 According to Theorem 3.1(e), it suffices to prove that the map $x \mapsto \langle Jx, \zeta \rangle$ is weakly continuous for given $\zeta \in X^*$. This holds because it is linear and strongly continuous.

5.49 It follows from Cor. 1.34 that the weak topology of L is the trace on L of the weak topology of X. Now L is weakly closed in X (being closed and convex), as is its unit ball. Thus, the unit ball in L is weakly compact, since the unit ball in X is weakly compact. Then Theorem 5.47 implies that L is reflexive.

5.52 This exercise is taken from Rudin [37].

5.53 It is a matter of appealing to Theorem 5.51, of course. The lower semicontinuity (and convexity) of f can be gleaned from the observation

$$f(x) = \sup_n \sum_{i=1}^n f_i(x_i).$$

See Exer. 9.6 for further analysis of the problem.

6.6 One applies Theorem 5.51.

6.7 There exists $v_* \in L^r(a,b)$ and a subsequence such that ${x_{i_j}}' \to v_*$ weakly in $L^r(a,b)$, by weak compactness. We may also suppose that $x_{i_j}(a) \to x_0$. Set

$$x_*(t) = x_0 + \int_0^t v_*(s)\, ds.$$

Then $x_* \in AC^r[a,b]$ and $x_{i_j}(t) \to x_*(t)$ for each t (why?). Using Hölder's inequality, prove that the functions x_{i_j} are equicontinuous, so that, by Ascoli's theorem, $x_{i_j} \to x_*$ uniformly for a further subsequence.

6.9 It consists of mappings of the type

$$x \mapsto \alpha \bullet x(0) + \int_0^1 \beta(t) \bullet x'(t)\, dt,$$

where $\alpha \in \mathbb{R}^n$ and $\beta \in L^q(0,1)$.

6.16 The unit ball in $L^\infty(\Omega)$ is weak* sequentially compact, by Theorem 3.21 and the separability of $L^1(\Omega)$.

6.18 It suffices to show that Φ is a closed convex (and hence weakly closed) subset of the set K defined in Prop. 6.17.

6.19 Prove that

$$\int_0^1 v_i(t)\,g(t)\,dt \;\to 0$$

for any smooth function g, using integration by parts. Deduce the same fact for $g \in L^\infty(0,1)$.

6.24 Let Δ_Γ and Δ_G be the effective domains of Γ and G. The effective domain of $F := \Gamma + G$ is $\Delta_F = \Delta_\Gamma \cap \Delta_G$. Let $\{\gamma_i\}$ and $\{g_j\}$ represent Γ and G as in Theorem 6.22. Then, for any compact subset V of $\mathbb{R}^n$, we have

$$\{x : F^{-1}(V) \neq \emptyset\} \;=\; \bigcup_{i\geqslant 1}\bigcap_{j\geqslant 1}\bigcup_{k,l\geqslant 1}\{x : |\gamma_k(x)| \leqslant i,\, \gamma_k(x) + g_\ell(x) \in V + j^{-1}B\}.$$

6.30 Let the functions γ_i generate G as in Theorem 6.22, and let $\Delta = \operatorname{dom} G$. We have

$$\{x \in \Delta : d_{G(x)}\big(u(x)\big) < r\} \;=\; \bigcup_{j\geqslant 1}\{x : |\gamma_j(x) - u(x)| < r\}$$
$$\{x \in \Delta : d_{G(x)}\big(u(x)\big) > r\} \;=\; \bigcup_{i\geqslant 1}\bigcap_{j\geqslant 1}\{x : |\gamma_j(x) - u(x)| > r + i^{-1}\}.$$

6.40 For the nonconvex case, consider Exer. 6.19.

6.42 We may suppose γ nonnegative. Use Gronwall's lemma to show that the sequence p_i is bounded in $C[a,b]$. Then $|p_i'(t)| \leqslant k(t)$ a.e., for a summable function k. It follows that the family p_i is equicontinuous. By Ascoli's theorem, a subsequence converges uniformly to a continuous function p. The criterion of Prop. 6.17 allows one to assume that p_i' converges weakly in $L^1(a,b)$ to a limit v. Passing to the limit in

$$p_i(t) \;=\; p_i(a) + \int_a^t p_i'(s)\,ds,$$

we find that $p \in \mathrm{AC}[a,b]$, where p_i' converges weakly in $L^1(a,b)$ to p'.

7.33 Use the lemma in the proof of Prop. 7.31.

7.40 This is an immediate consequence of Theorem 7.34 and Prop. 7.39.

8.3 Without loss of generality, assume $0 \in A$. Then, because A is open, it follows that

$$c_x > 0 \;\;\forall\, x \in \partial A,\;\; \langle \zeta_x, u \rangle < c_x \;\;\forall\, u \in A.$$

We deduce

$$A \subset \bigcap_{x\in\partial A}\{u : \langle \zeta_x, u \rangle < c_x\}.$$

We now prove the opposite inclusion, which reveals that A is the intersection of convex sets, and is therefore convex itself. Let $y \notin A$. Then there exists a first t in $(0,1]$ such that $x := ty$ lies in ∂A. We have

$$\langle \zeta_x, ty \rangle = c_x > 0 \implies \langle \zeta_x, y \rangle \geqslant c_x,$$

and it follows that $y \notin \{u : \langle \zeta_x, u \rangle < c_x\}$. This argument adapts easily to the case in which A is closed and $0 \in \operatorname{int} A$, using closed halfspaces rather than open ones, and observing that the t above satisfies $t < 1$.

8.4 Use Baire's theorem.

8.5 Hint: consider the truncations of α_n, and uniform boundedness.

8.6 For the "if" part: when the stated condition holds, then, for every $\varepsilon > 0$, the function

$$u \mapsto -\langle \zeta, u \rangle + \varepsilon |u - x|$$

has a local minimum over S at $u = x$, whence $\zeta \in \varepsilon B + N_S(x)$ by Prop. 4.12.

8.9
(a) The map $(x,u) \mapsto x+u$ is weak* continuous, and $\Sigma \times \Delta$ is weak* compact.

(b) Observe that the set consists of all convex combinations $\sum_1^n t_i \sigma_i$, where $\sigma_i \in \Sigma_i$.

8.11 If (c) holds, then the complement of $N(\zeta)$ contains a ball $x + rB$. If $\zeta(rB)$ is bounded, (a) holds. Otherwise, $\zeta(rB) = \mathbb{R}$. Then $x + rB$ intersects $N(\zeta)$: contradiction. The result is from Rudin [38, Theorem 1.18].

8.16 Consider Example 3.11.

8.17 We may assume $f(0) = 0$. Pick $r > 0$ so that $f(x) \geqslant 1$ on the set $\|x\| = r$. Then, for any u with $\|u\| \geqslant r$, we have

$$1 \leqslant f\left(\frac{r}{\|u\|}u\right) \leqslant \left(1 - \frac{r}{\|u\|}\right) f(0) + \frac{r}{\|u\|} f(u) = \frac{r}{\|u\|} f(u).$$

8.19 Let v be a unit vector. We have $f'(x;v) = \inf_{t>0} [f(x+tv) - f(x)]/t$. Taking $t = |x|+1$ and applying the bounds on f leads to $f'(x;v) \leqslant c + d|x|$, for certain constants c and d.

8.20 One can argue via support functions, by showing that

$$f'(x;v) \leqslant \max\{g_i'(x) \bullet v : i \in I(x)\}.$$

8.21 Take f to be the indicator function of the set C of Exer. 2.49.

8.22 Show that f is convex (Exer. 8.7); use Prop. 4.12. (This exercise is purloined from [7].)

8.24 For (b): For fixed u sufficiently close to x ($u \neq x$), the function $g(t) = f(x + t(u-x))$ is differentiable for t in a neighborhood of $[0,1]$. We may apply the (one dimensional) mean value theorem to g in order to write $f(u) - f(x) = \langle f_G'(w), u - x \rangle$ for some $w \in (x,u)$. Then

$$\begin{aligned}\|f(u) - f(x) - \langle f_G'(x), u - x \rangle\|_Y &= \|\langle f_G'(w) - f_G'(x), u - x \rangle\|_Y \\ &\leqslant \|f_G'(w) - f_G'(x)\|_* \|u - x\|,\end{aligned}$$

which implies that f is differentiable at x.

8.25 Let $f(x) = \|x\|$. If $x \neq 0$ and $\zeta \in \partial f(x)$, then $\langle \zeta, x \rangle = \|x\|$ and $\|\zeta\|_* = 1$ (see Exer. 4.2). The strict convexity of B_* implies that there cannot be more than one such ζ.

8.27 Concerning the second part: when $X = \mathbb{R}$, the result is easy to prove. Build on this case to show that f is convex when restricted to any line.

8.28
(a) Hint: TB_X contains a neighborhood of 0.

(c) Use Theorem 5.19. ($T(X)$ may not be closed, so separation does not apply.)

(d) Recall that the function $g(y) = \|y\|_Y$ is convex, and that if $\zeta \in \partial g(y)$, where $y \neq 0$, then $\|\zeta\|_* = 1$ (see Exer. 4.2).

8.30 Invoke Cor. 4.23.

8.31

a) Show that f^* is finite-valued, and (with the help of Exer. 4.17) that $\partial f^*(x)$ is a singleton for every x. Then appeal to Prop. 4.16.

b) If g^* fails to be strictly convex, then (by Exer. 4.17) there exist distinct points x and y and $\zeta \in \partial g^*(x) \cap \partial g^*(y)$. Then $x, y \in \partial g(\zeta)$, a contradiction.

8.32 Apply Theorem 5.19 to f.

8.33 One may first reduce to the case in which $\|T\| < 1$, and then argue as in Rudin [38, Theorem 10.7].

8.34 The conclusions can be obtained from Theorem 5.32, with $F(x,r) = (Tx + rg(x), r)$. There is a minor technical point to deal with, due to the fact that the distance appearing in the conclusion of the theorem may not be attained.

8.35 One approach involves considering

$$\varphi(x,q) = \min_{|v| \leqslant 1} f'(x;v) - q \bullet v,$$

showing that $q \in \partial f(x)$ if and only if $\varphi(x,q) = 0$, and using Prop. 6.25.

8.36 $\Gamma(x)$ consists of the points q satisfying $f^*(x) = x \bullet q - f(q)$, and is nonempty.

8.38 Hint: there is a point $y \notin S$ at which $\partial_P d_S(y) \neq \emptyset$.

8.39

(c) We have $\varphi(x) = \sup_{u \in S} s \bullet x - |s|^2/2$, implying convexity.

(f) If $y \in \partial g(x)$, then

$$x \in \partial \varphi(y) = \{\varphi'_G(y)\} = \{y - (y - s_y)\} = \{s_y\}$$

(a point in S), by the derivative calculation in part (b), and since $\partial \varphi(y)$ reduces to the singleton $\{\varphi'_G(y)\}$ by Cor. 4.4.

(g) Invoke Prop. 5.21 to obtain $\overline{A} \supset \operatorname{dom} g$.

8.40 This can be proved with the help of Theorem 5.19; see [13, Theorem 7.6.2].

8.42 For the last part: we have that c_0^{**} is isometric to ℓ^∞. The sequence $(1, 1, \dots) \in \ell^\infty$ corresponds to an element of c_0^{**} that does not lie in Jc_0.

8.43

(a) The countable base can be assumed to consist of canonical open sets for the weak topology; the collection of the associated ζ_n is a countable set. Arguing as in part (c) of Theorem 3.1, we see with the help of Exer. 2.41 that any ζ is a finite linear combination of these.

(b) This follows from Exer. 8.4, since X^* is a Banach space.

8.45 The main difficulty is related to Exer. 6.40.

8.47 Argue as in the proof of Theorem 5.47 to show that the topologies $\sigma(X^*, X)$ and $\sigma(X^*, X^{**})$ coincide.

8.48 For (a): First, invoke strict separation to find $\theta \in X^{**}$ such that $\langle \theta, \zeta \rangle < 0 \leqslant \langle \theta, \sigma \rangle$. Adjust θ in order to have both inequalities strict. Then call upon Goldstine's lemma to conclude. Part (c) is a special case of a classical theorem of Krein-Šmulian; no simple proof is known to us. See Rudin [38, Exer. 4.21] for a proof outline and references.

8.49 For the last assertion, invoke Exer. 7.36.

8.50 There are no points of differentiability.

8.51 In either case, one invokes the minimax theorem 4.36. When Σ is bounded, the required compactness hypothesis is provided in part with the help of Exer. 8.47.

8.53 One could consider $e^{-\|x\|}$ on $L^1(0,1)$.

8.57 K is compact in either case, with respect to the appropriate weak topology. When $\sigma(Y^*,Y)$ is involved, separation calls upon Prop. 3.12. Note that the closure operation in $\overline{\text{co}}\,E$ is to be interpreted for the relevant topology.

Part II. Optimization and nonsmooth analysis.

The subject of optimization boasts an enormous literature. For an introduction to the topic, the reader could do worse than the book of Boyd and Vandenberghe [7]. The calculus of generalized gradients, along with the use of proximal normals, was introduced in the author's thesis [11]. His subsequent book on nonsmooth analysis [13] was the first in the field, and has been widely cited. More recent sources include the monographs of Clarke, Ledyaev, Stern and Wolenski [18], and that of Rockafellar and Wets [35], which contain a variety of applications of the subject, as well as detailed references.

9.2 Both x_2 and x_4 fail to admit multipliers, while x_3 is not admissible; there appears to be no reason to veto x_1.

9.7 If the functions involved are differentiable, and if $x_* \in \operatorname{int} S$, then we obtain the classical stationarity conclusion.

9.11 One must show, using $(\gamma,\lambda) \in -\partial V(0,0)$, that $\gamma \geqslant 0$, and that the complementary slackness and minimization conditions hold (for the solution x_*).

10.26
(a) If $F(x) < 0$, the optimality of x_* is contradicted.

(b) Apply Danskin's theorem, and show that the set $\{\varphi'_x(x_*,y,z) : y \in Y^*, z \in \{0,1\}\}$ is closed; then invoke Carathéodory's theorem.

(d) Observe that in either case, the necessary conditions amount to asserting that 0 belongs to the set $\operatorname{co}\{f'(x_*), e'_i(x_*)\ (i \in I(x_*)\}$, where $I(x_*)$ is the set of indices corresponding to the *active* constraints.

10.29 For convexity, one may show that $\partial_C f$ is not monotone (Exer. 8.27); for example:

$$\langle \nabla f(1,0) - (2,1), (1,0) - (0,0) \rangle < 0.$$

For the regularity: $f'(0,0;1,1) = 2$, but $f^\circ(0,0;1,1) \geqslant (1,1) \bullet (2,1) = 3$.

10.30 Use the gradient formula to show that $\partial_C g(x) = \{0\}$ at every point x in Ω, and then conclude that g is constant with the help of Theorem 10.17.

10.43 Reduce to the case $0 \in \operatorname{int} S$, and consider the gauge of S.

10.48 To derive the multiplier rule for (Q), introduce an extra variable y, together with the constraints $\varphi(x) - y = 0$ and $(x,y) \in S \times \Phi$; invoke Theorem 10.47.

11.17 Write $\langle \lambda_i, F \rangle = \langle \lambda, F \rangle + \langle \lambda_i - \lambda, F \rangle$ and invoke Theorem 11.16, along with Prop. 11.12 and the estimate of Prop. 10.5.

11.21 Take $X = \mathbb{R}^2$, $f(u) = |u|$, $x = 0$, $W = \{u : |u| = 1\}$, $\rho = 1/2$.

13.1 The *Courant-Fisher formulas* extend the result by characterizing *all* the eigenvalues of M; there are n of these, if each eigenvalue is counted according to its multiplicity. Would the reader care to guess what optimization problem characterizes the *second* eigenvalue λ_2, when λ_2 is strictly greater than λ_1? (That is, when λ_1 is of multiplicity one.) [Hint: an eigenvector for λ_2 is necessarily orthogonal to one for λ_1.]

13.3 Use Exer. 9.7.

13.4 It follows that Λ is continuous in (t,v), and that the cost integral is well defined, as well as convex in $v(\cdot)$. The necessity of the stated condition can be derived from Theorem 9.8 (together with Theorem 6.32). The necessity results from a direct argument.

13.7 For (c), use Prop. 4.6 and Theorem 2.34.

13.8 Invoke Exer. 13.7 with $f(x) = I_S(x)$ and $g(x) = \|x - y\|$.

13.9 The dual problem consists of maximizing over $(0,\infty) \times \mathbb{R}^n$ the function

$$(\gamma, \lambda) \mapsto -E_0\,\gamma - \sum_k \lambda_k c_k - \int_{-\pi}^{\pi} \Big[\,\Big|\sum_k \lambda_k \cos(kt)\Big| - 1\Big]_+^2 \,dt\,/(4\gamma).$$

13.10 Use the gradient formula to calculate $\partial_C f(0)$.

13.11
For S_1: $T = S_1$, $N = \{0\}$, $N^L = S_1$, $N^C = \mathbb{R}^2$, $T^C = \{0\}$.

For S_2 (convex): $T = T^C = S_2$, $N = N^L = N^C = \{y \leqslant -|x|/2\}$.

For S_3: $T = S_3$, $N = \{0\}$, $N^L = \{y = |x|/2\}$, $N^C = \{y \geqslant |x|/2\}$, $T^C = -S_2$.

For S_4: $T = \mathbb{R}^2$, $N = \{0\}$, $N^L = N^C = \mathbb{R} \times \{0\}$, $T^C = \{0\} \times \mathbb{R}$.

13.16 Prove the implication first in the case $(\theta, \zeta) \in \partial_P f(u,v)$, by examining the proximal subgradient inequality. Then consider $(\theta, \zeta) \in \partial_L f(u,v)$.

13.17 For the last part: we have $g(u) \leqslant g(x) + K|u - x|$ for u near x, which implies that $g = f$ locally near x.

13.18 Hint: use Cor. 11.7.

13.19 This follows from Prop. 7.39.

13.22 For (b): Use Theorem 11.38.

13.23 Suggestion: use Theorem 10.19 to estimate $\partial f(1) = \partial_C f(1)$, with $F(\alpha) = (\alpha, v_*/\alpha)$ and $g(\alpha, w) = \Lambda(w)\,\alpha$; Exer. 10.21 may be useful too.

13.24 For (b): It follows from the definition of $\partial_L f$ that the set in question is given by

$$\bigcap_{j \geqslant 1} \bigcup_{i \geqslant 1} \big\{t : 0 \in \theta_i\big(t, u(t) + j^{-1}B, V + j^{-1}B\big)\big\}.$$

For each i, j, the last set above is the domain of the multifunction

$$t \mapsto \big\{(x,\zeta) \in \big(u(t) + j^{-1}B\big) \times \big(V + j^{-1}B\big) : \theta_i(t,x,\zeta) = 0\big\},$$

which is measurable by Prop. 6.25. The measurability of the multifunction $t \mapsto \partial_C f(t, u(t))$ follows from Cor. 6.28.

13.30 Time reversal is involved in the proof; details are given in [18, Prop. 6.5].

13.33 For the first part, we argue by contradiction. If the assertion fails, there exist sequences x_i, α_i, and $\varepsilon_i > 0$ converging to $x_*, F(x_*)$, and 0, with $F(x_i) \neq \alpha_i$, such that, for some ζ_i belonging to $\partial_P |F(x_i) - \alpha_i|$, we have $|\zeta_i| < \varepsilon_i$. By Exers. 7.33 and 13.32, we have $\zeta_i \in v_i^* \partial F(x_i)$ for a certain unit vector v_i. Taking subsequences, we derive $0 \in v^* \partial F(x_*)$ for some unit vector v; this contradicts the nonsingularity.

13.34 A detailed proof is given in [13, p. 253].

Part III. Calculus of variations.

This is an old subject, and volumes have been written about it. The reader will have no difficulty in finding parallel material. For our part, we have liked and benefited from the little books of Bliss [5] and of Gelfand and Fomin [26], the big book of Morrey [31], and the ineffable book of L. C. Young [41]. We also take pleasure in citing the works of Cesari [9], Ewing [23], Goldstine [27], and Troutman [39].

The necessary conditions of Chapter 18, with those of Chapter 25, are the end product of a thirty-year quest by the author (and certain colleagues). The results in final form were first described in the monograph [16], which contains a detailed discussion of related work.

14.3
(a) We expect the shortest curve joining two points to be a line segment.

(c) We find $x(t) = (e^t - e^{-t})/(e - e^{-1})$.

14.15 Consider the Legendre condition.

14.16 Let $T < \tau$, the nearest conjugate point to 0; apply Theorem 14.12.

14.17 The Jacobi equation is $u'' + u = 0$, for which $u(t) = \sin t$ is a suitable function for locating conjugate points; we find $\tau = \pi$. As observed previously, local and global minima coincide for this problem. For $T < \pi$, the proposition therefore follows from Theorem 14.12; the case $T = \pi$ is obtained by approximation.

14.20 For the modified problem: $x_*(t) = t^2/2 - 4t$.

14.23 Apply Theorem 14.21, and show that the abnormal case $\eta = 0$ can be excluded. Then the solution satisfies $x'' = \lambda x$. It follows that $\lambda < 0$, which leads to $x(t) = \pm \sin kt$, with k a positive integer. This gives $J(x) = k^2 \pi/2$, whence $k = 1$ and $x_*(t) = \pm \sin t$.

14.24 The possibilities are $x(t) = \pm(\sin kt)/k$, for a positive integer k. The corresponding cost is $\pi[2k^2]^{-1}$. This admits no minimum relative to k, so no solution exists. There is, however, a sequence x_i of admissible sawtooth functions converging uniformly to 0, whose associated costs $J(x_i)$ converge to 0. This is clearly the (unattained) infimum.

15.6
(a) If $u'(\tau) = 0$, then, since $u(\tau) = 0$, we have $u \equiv 0$ by the uniqueness theorem for linear differential equations; absurd. Use integration by parts together with the Jacobi equation to obtain the stated equality.

(b) By Theorem 15.5, u (extended) is continuously differentiable on (a,b); yet u has a corner at τ: contradiction.

15.13
(a) The Euler equation for Λ_+ (see part (b)), together with the boundary conditions and the isoperimetric constraint, identify the candidate $x_*(t) = t + 2(1/\pi - 1)\sin t$.

(b) It follows from Theorem 15.9 that for any $x \in \mathrm{Lip}[0,\pi]$ satisfying the boundary conditions, we have

$$\int_0^\pi x'^2(t)\,dt \;+\lambda\int_0^\pi (\sin t)x(t)\;dt \;\geqslant\; \int_0^\pi x_*'^2(t)\,dt + \lambda\int_0^\pi (\sin t)x_*(t)\;dt\,.$$

If x satisfies in addition

$$\int_0^\pi (\sin t)x(t)\;dt \;=1=\; \int_0^\pi (\sin t)x_*(t)\;dt\,,$$

then we deduce $\int_0^\pi x'^2(t)\,dt \;\geqslant\; \int_0^\pi x_*'^2(t)\,dt$. Thus, x_* solves (Q).

16.3 Hints for the five parts: coercivity, weak compactness and Ascoli, dominated convergence, Fatou, weak lower semicontinuity.

16.10 Show that $J(x)$ is well defined for any arc x, and note that a minimizing sequence x_i exists. Use Hölder's inequality to show that the sequence x_i' is bounded in $L^r[a,b]$. Define an appropriate (bounded) set Q for purposes of invoking the integral semicontinuity theorem 6.38.

16.22
(a) Use the inequality $e^y \geqslant 1+y$ in analyzing a minimizing sequence.

(b) Show that the Lagrangian is strictly convex.

(d) Invoke Theorem 16.18.

17.11 Show first that Theorem 17.9 applies, and that the necessary conditions hold normally.

17.12
(a) In applying the direct method, note that the isoperimetric constraint provides a lower bound on the second term in $I(x)$.

(b) Show that Theorem 17.9 can be called upon.

(c) Integrating by parts leads to $\lambda = I(u)$.

18.5 Derive $p(t) \in \partial_v \Lambda(t, x_*(t), x_*'(t))$ a.e. (see Exer. 13.16). Then show that the proof of Theorem 15.5 can be adapted.

18.10
(a) A solution $x_* \in \mathrm{AC}[0,T]$ exists by Tonelli's theorem. It is unique by strict convexity, and x_* is Lipschitz by Theorem 16.18. Exer. 18.5 shows that $x_* \in C^1[0,T]$.

(b) Apply Theorem 18.8, with the costate

$$p(t) = \begin{cases} -t+\sqrt{2} & \text{if } 0 \leqslant t \leqslant \sqrt{2} \\ 0 & \text{if } \sqrt{2} < t < T-\sqrt{2} \\ t-(T-\sqrt{2}) & \text{if } T-\sqrt{2} \leqslant t \leqslant T. \end{cases}$$

(e) Show first that the only possible turnpike value is $x_* = 0$. But then (by Cor. 16.19) we would have $|x_*(t)| + |x'(t)|^2/2 = 0\ \forall t$, whence $x_* \equiv 0$ and $A = 0 = B$ necessarily. Conclude by showing that when $A = 0 = B$, the zero arc *fails* to be a minimizer.

19.5 Show that, for any arc x, we have $(d/dt)\,u(t,x(t)) \leqslant \Lambda(x(t), x'(t))$, and integrate.

19.8 For (b): consider Cor. 11.46. For (d): observe that in a neighborhood of $(1,0)$, we have $u(t,x) = |x|$.

19.9 If (θ,ζ) belongs to $\partial_P \min(u,v)(t,x)$ at a point (t,x) where $u = v$, then $(\theta,\zeta) \in \partial_P u(t,x)$.

19.13 We solve the problem

$$\min\ |x(0)|^2 + \int_0^\tau |x'(t)|^2/4\,dt\,,\ \ x(\tau) = \beta\,,$$

and this leads to the function $u_*(t,x) = |x|^2/(1+4t)$.

19.14 The minimization problem yielding $u(t,x)$ is given by

$$\text{minimize}\ \ |y(0)| + \int_0^t \{\,|y'(s)|^2/4 + 1\,\}\,ds\,,\ \ y(t) = x\,.$$

The problem may be solved deductively (by Theorem 18.1) or inductively (Theorem 18.8) to yield

$$u_*(t,x) = \begin{cases} |x|^2/(4t) + t & \text{if } |x| < 2t \\ |x| & \text{if } |x| \geqslant 2t\,. \end{cases}$$

19.15 We know that $u_*(\tau,\beta)$ is the minimum of the cost function

$$J(x) = \ell\big(x(0)\big) + \int_0^\tau \Lambda\big(x'(t)\big)\,dt$$

over the arcs x satisfying $x(\tau) = \beta$. Solutions of this problem are Lipschitz (Theorem 16.18). Because Λ is strictly convex (see Exer. 8.31), it follows from the necessary conditions of Theorem 18.1 that the solutions are affine. The cost corresponding to an admissible affine function is given by $\ell(\alpha) + \tau\Lambda((\beta-\alpha)/\tau)$, where $\alpha = x(0)$, whence the stated formula.

19.21 Hint: See the last step in the proof of Theorem 16.18.

20.9
(a) Use Green's theorem 20.7.

(b) Call upon Theorems 20.6 and 20.8.

(c) If u_1 and u_2 are both weak solutions of (D), then both minimize the Dirichlet functional. Because that functional is strictly convex in Du, this implies that Du_1 and Du_2 agree almost everywhere. But then $u_1 = u_2$ by Exer. 10.30.

20.16 Invoke the comparison principle with the constant function $M = \max\{\,u_*(x) : x \in \Gamma\,\}$.

21.1
(a) $x_*(t) = -t/2 + 1/2$.

(b) $\Lambda(t,x,v)$ is convex in (x,v), so that x_* is a global minimizer relative to Lip$[1,3]$, as well as $C^2[1,3]$.

(c) $\Lambda_{vv} < 0$ for $t < 0$, which precludes a local minimum by Legendre's necessary condition.

21.2
(e) This follows from the Weierstrass condition. Alternatively, observe that there exist admissible Lipschitz arcs x whose derivative is 0 or -1 a.e., so that $J(x)$ is 0. Approximation (see Exer. 21.13) implies the existence of a smooth admissible y such that $J(y) < J(x_*)$.

21.3
(a) $x_*(t) = -t^2 - 2 + ce^t + de^{-t}$, where $c + d = 2$, $ce + d/e = 4$.

(c) $x_*(t)$ has the same form, but now $c+d=2$, $ce-d/e=2$ (by the transversality condition).

21.4 For the last part, a simple verification function can easily be found by inspection.

21.5 For (a), prove that there exist unique numbers $c<0$ and $\lambda>0$ so that the circle

$$(t-1/2)^2+(x-c)^2 = \lambda^2$$

defines an admissible function x_*; it satisfies $\lambda\, x'/\sqrt{1+x'^2} = -t+1/2$. We obtain (c) by simply specializing the conclusion of (b) to those x satisfying the isoperimetric constraint.

21.6 The problem is taken from [26]. If a smooth solution is simply assumed to exist, it is natural to apply the multiplier rule of Theorem 14.21 in $C^2[0,b]$ to the problem

$$\min \int_0^b x(t)\sqrt{1+x'(t)^2}\,dt \ : \int_0^b x(t)\,dt = S,\ x(0)=A,\ x(b)=0.$$

The Erdmann condition for $x\sqrt{1+v^2}+\lambda x$ yields $1+\lambda\sqrt{1+x'(t)^2}=0$, whence x' is constant. Thus, the solution corresponds to the line $t/c+x/A=1$, where $cA=2S$.

However, one's conscience tends to require a proof that this is truly optimal in some sense. If the problem is restricted to curves which can be expressed as $t(x)$, then the problem becomes

$$\min \int_0^A x\sqrt{1+t'(x)^2}\,dx \ : \int_0^A t(x)\,dx = S,\ t(0)=b,\ t(A)=0 \ (b \text{ free}).$$

This reformulation has the advantage of being convex. The necessary conditions lead again to the affine candidate, but now its optimality follows rigorously (even in the presence of the implicit constraint $t(x)\geqslant 0$, but with the restricted class), by an argument similar to that of Exer. 21.5.

The best approach uses verification functions. Conjecturing that the solution is affine, we calculate the (expected) optimal cost

$$V(S,A) = \sqrt{4S^2+A^4}/2.$$

The function $V(s,x)=\sqrt{4s^2+x^4}/2$ satisfies, for $x\geqslant 0$, the inequality

$$V_s x - V_x v \leqslant x\sqrt{1+v^2},$$

as one can see by calculating the derivatives and squaring both sides. Now let x be admissible. Then we may write

$$V_s\Big(\int_t^b x(\tau)\,d\tau,\, x(t)\Big)x(t) - V_x\Big(\int_t^b x(\tau)\,d\tau,\, x(t)\Big)\,x'(t) \leqslant x(t)\sqrt{1+x'(t)^2}$$

and integrate to get

$$J(x) \geqslant V\Big(\int_0^b x(\tau)\,d\tau,\, x(0)\Big) - V(0,0) = V(S,A),$$

proving that the affine arc is indeed optimal.

21.7 (b) $\exp(x+v)$ (c) xe^v

21.8 The corresponding Lagrangian is $v^2/2+(\cos^3 x)/3-2\cos x$. Reason as in Example 16.21.

21.9 For (b): If there were a conjugate point in (a,b), then x_* would not be a minimizer; but it is, by convexity.

21.10 The arc $x_*(t)=t$ is an admissible extremal; the Lagrangian is convex.

21.11 In this reformulation, the Lagrangian is convex.

21.12 The Euler equation identifies the extremal $x_*(t) = \ln(1+t)/\ln 2$. The convexity of Λ implies that x_* is the global minimizer (Theorem 18.8). There cannot be any local minimizers or maximizers.

21.13 For (c), write

$$\varphi(b) = h(a) + \int_a^b \big\{ (p-f) + (f-h') + h' \big\}\, dt\,,$$

and consider in light of this how to choose c so that g agrees with h at a and b.

(f) Apply (e) to the Lagrangian

$$\widetilde{\Lambda}(t,y,w) \;=\; \Lambda(t, x_*(t)+y, x_*'(t)+w).$$

21.14 Build upon Exer. 14.17, and use Exer. 21.13.

21.15 To reason deductively, prove first that a solution x_* exists, by the direct method. Then Cor. 16.19 implies that x_* is Lipschitz, and that $x_*^2 + x_*'^2 = c^2$ for some constant $c > 0$. The state constraint must be active at some point, whence $c \geqslant 1$. Show that $x_*(t) > 0$ for all $t > 0$ sufficiently near 0. Until $x_* = 1$, the Euler equation holds, and implies that $x_*(t) = c \sin t$. After x_* reaches the value 1, it remains there (this is evidently optimal), whence $c = 1$. This identifies the unique solution x_*.

21.16
(a) The Jacobi equation has the solution $u = (7t+1)^{2/3}$, so that Cor. 14.13 applies.

(b) Prove first that a solution x in $\mathrm{AC}[0,1]$ exists. By Cor. 16.19 we have

$$\big(1 + x(t)\big)\, x'(t)^2 \;=\; c \text{ a.e.}$$

Show that $c > 0$, and that there must be $\varepsilon > 0$ arbitrarily small such that $x(\varepsilon) > 0$ and $x' > 0$ on a set of positive measure near ε. Using the fact that the Euler equation holds on intervals where $x(t) > 0$, deduce that x' is continuous. Show that $x(t) > 0$ for $t > 0$, and go on to prove that $x = x_*$.

21.17 For (a), let

$$\Lambda(t,x,v) = x^4/4 + b\,x^3/3 + c\,x^2/2 + d\sin(t)\,x + v^2/2\,,$$

and deduce the existence of a minimum $x_* \in \mathrm{Lip}[0,1]$ (use Cor. 16.16). Then x_* belongs to $C^\infty[0,1]$ by Theorems 15.5 and 15.7, and the Euler equation may be written in differentiated form.

(b) When $b^2 \leqslant 3c$, Λ is convex in (x,v) (as well as strictly convex in v). It follows that the minimum x_* is unique. But a solution of (1) is a global minimizer for (P), by convexity.

(c) Hint: If the zero function has a conjugate point in $(0,T)$, it cannot be the solution of (P).

21.18 For (a) and (c), the weak closure theorem 6.39 is useful in combination with the direct method.

21.23 For (b): $t^4 - 2t^3 + t^2 + t$.

21.24 Prove that there is a solution x_* of the problem

$$\text{minimize } \int_0^1 \big\{ |x'(t)|^2/2 + f(x(t)) \big\}\, dt \;:\; x \in \mathrm{AC}[0,1]\,,\, x(0) = A\,,\; x(1) = B\,.$$

Then x_* is Lipschitz by Theorem 16.18, and C^1 by Theorem 15.5. The Euler inclusion of Theorem 18.1 implies that $x_*' \in \mathrm{AC}[0,1]$ and that x_* satisfies the required differential inclusion.

21.25 One may calculate the value function $\varphi(t,x) = x^2 \cot t$.

21.26 See [39, pp. 34, 190, 320] for the full statement of the problem, and a different approach to solving it.

21.30 Details and references, as well as for Exer. 21.31, are given in Clarke [13].

21.31
(a) Hint: Fix $r \in (2, R_*)$; write Young's inequality: $|q' \bullet y| \leqslant |q'|^r/r + |y|^{r_*}/r_*$.

(b) Use Theorem 16.18.

(c) Call upon Theorem 18.1.

(f) Show that if T is not the minimal period, then $(y(t/k), q(t/k))$ (for some integer $k > 1$) is strictly better than (y,q) for the problem (P), a contradiction.

21.32
(a) Taking $c = 0, d \equiv 1$, the inequality in Hypothesis 18.11 follows from the fact that any (ζ, ψ) in $\partial_P \Lambda(x,v)$ satisfies $\psi + \zeta = 1$ (see Exer. 7.27).

(b) If x_* is a strong local minimizer, then there is an arc p satisfying the conclusions of Theorem 18.13. The Euler inclusion for x_* gives $p' + p = 1$, and transversality provides $p(1) = 0$; it follows that $p(t) = 1 - e^{1-t}$. The Weierstrass condition reads

$$\sqrt{|v|} + v \geqslant \langle p(t), v \rangle \ \ \forall v \in \mathbb{R},$$

which forces $p(t) = 1$: contradiction.

(c) It suffices to consider arcs x for which $x(t) \leqslant 0$ and $x'(t) \leqslant 0$. For any such admissible arc x with $|x'(t)| \leqslant e^{-2}$ we have $|x(t)| \leqslant e^{2(t-1)}$, whence

$$|x(t) - x'(t)| \leqslant e^{2(t-1)}.$$

We calculate (the reader will discern a verification function at work)

$$\begin{aligned} \sqrt{|x(t) - x'(t)|} + x'(t) &\geqslant |x(t) - x'(t)|\, e^{1-t} + x'(t) \\ &\geqslant (x(t) - x'(t))\, e^{1-t} + x'(t) = \frac{d}{dt}\left\{ x(t)(1 - e^{(1-t)}\right\}. \end{aligned}$$

The result follows upon integrating. Note that the classical methods to prove the presence of a weak local minimizer do not apply here.

21.33
(a) The facts stated in Exer. 4.27 are relevant here.

(e) With $h = -1$, it follows that $x_* \equiv 0$, whence $u(\tau, 0) = 2\tau$: contradiction.

21.34 Consider the following closed subspace X of $L^p(\Omega)^{n+1}$:

$$X = \left\{ \big(u, Du\big) : u \in W_0^{1,p}(\Omega) \right\}.$$

Show that ζ corresponds to an element of the dual of X. For the second part, consider pairs of the form $(f_0 + \operatorname{div} p, f + p)$.

21.35 Because weakly convergent subsequences are bounded, there is a subsequence such that u_{i_j} and Du_{i_j} converge weakly in L^p; the convergence must be to u_* and Du_* respectively; apply Theorem 6.39.

21.36 If such an operator T exists, then there is a number K such that, for every $u \in C(\overline{\Omega})$:

$$\| u|_{\partial\Omega} \|_{L^p(\partial\Omega)} \leqslant K \| u \|_{L^p(\Omega)}.$$

For any u for which the left side is positive, we can modify u (in Ω only, not on the boundary) so that the right side is arbitrarily small: contradiction.

21.37 For (b): The weak Euler equation is

$$\int_\Omega \big\{ \langle Du, D\psi \rangle + (u-\theta)\psi \big\} \, dx\, dy = 0 \;\; \forall\, \psi \in C^2(\overline{\Omega}).$$

The classical divergence theorem implies

$$\int_\Omega \big\{ \langle Du, D\psi \rangle + \psi \operatorname{div} Du \big\} \, dx\, dy = \oint_\Gamma \psi \, Du \bullet \nu \, d\gamma.$$

Together, these lead to the conclusion.

21.38 Existence and uniqueness follows from Cor. 20.24. For the lower bound on u, recall that u is a minimizer for the Lagrangian $|Du|^2$; use the comparison principle (Theorem 20.15) (with u and a certain constant function).

21.40 For (a): Use Cor. 4.7 and Prop. 4.14.

Part IV. Optimal control.

The books that have influenced us the most on this topic are those of Pontryagin et al. [33], Young [41], Lee and Markus [30], and Hestenes [28]. Among other favorites of ours are Vinter [40], Fleming and Rishel [24], and Bardi and Capuzzo-Dolcetta [4]. Our own earlier book [13] has even taught us a thing or two, given the passage of time.

The extended maximum principle first appeared in the author's thesis [11], which also initiated the study of necessary conditions for differential inclusions. The monograph [16] contains a detailed discussion of the body of related work.

22.7 The new component q of the costate satisfies $-q' = H_t^{\eta}$, $q(b) = 0$. Combine this with the constancy of the augmented Hamiltonian $q + H^{\eta}$.

22.27 The proof of Cor. 22.3 can be adapted.

22.29 The optimal control in the special case referred to is given by

$$u(t) = 1 \;\text{ for }\; t \in (0, 1/3) \cup (2/3, 1), \;\text{ and }\; 0 \;\text{ elsewhere}.$$

23.6 For a uniform partition π of the underlying interval, as in the proof of Theorem 12.3, define a piecewise affine (original) trajectory x by taking $x' = 1$ on $[t_i, t_{i+1})$ if $x(t_i) \leqslant y(t_i)$, and $x' = -1$ if $x(t_i) > y(t_i)$. Then take the mesh size of the partition sufficiently small.

23.7 One may adapt the proof of Theorem 23.2.

23.12
(a) To apply the existence theorem 23.11, one may consider (without changing the problem) that the running cost integrand is given by the *convex* function

$$\Lambda(u) = \begin{cases} 0 & \text{if } u \leqslant 0 \\ u^3 & \text{if } u \geqslant 0. \end{cases}$$

(b) The abnormal case ($\eta = 0$) is easily excluded. The costate p is constant, and the function $u \mapsto u^3$ is strictly convex on the interval $[0,2]$. Thus the maximum of $u \mapsto pu - u^3$ over the control set is attained at a unique point u_*.

23.16 For (b), show that the function

$$f(\lambda) = \Lambda\big(t, x, (1-\lambda)u + \lambda v\big) - (c/2)|(1-\lambda)u + \lambda v|^2$$

is convex by calculating f'', then write $f'(1) \geqslant f'(0)$.

24.4 With the help of the extended maximum principle, one identifies the following candidate (x,u) and corresponding costate p:

$$\big(x(t), u(t)\big) = \begin{cases} (t,1) & \text{if } 0 \leqslant t < 1/2 \\ (1/2, 0) & \text{if } 1/2 \leqslant t < 3/2 \\ (2t - 5/2, 2) & \text{if } 3/2 \leqslant t < 5/2 \\ (t,1) & \text{if } 5/2 \leqslant t \leqslant 3 \end{cases} \qquad p'(t) = \begin{cases} 0 & \text{if } 0 \leqslant t < 1/2 \\ -1 & \text{if } 1/2 \leqslant t < 1 \\ +1 & \text{if } 1 \leqslant t < 2 \\ -1 & \text{if } 2 \leqslant t < 5/2 \\ 0 & \text{if } 5/2 \leqslant t \leqslant 3 \end{cases}$$

with $p(t) = 0$ precisely when $t \in [0,1/2] \cup \{3/2\} \cup [5/2,3]$. We may apply Theorem 24.1 to deduce that the process is optimal.

24.8 Show that the system $(x^0 + T(x), F_+)$ is both weakly decreasing and strongly increasing on Ω, by the definition of T; this uses the fact that minimal-time trajectories exist.

26.1 The turnpike value is $(\delta + \sqrt{\delta^2 + 4}\,)/2$. Note that this increases with δ: it's weedier in inflationary times.

26.2
(b) Show that σ must attain 0 at a first point $\tau \in (0,T)$; note that $\sigma' = p$, and deduce $\sigma'(\tau) < 0$; show that σ becomes and remains negative after τ.

(c) Observe that in $[0,\tau]$ we have $x(t) = e^t - 1$, and that in $[\tau, T]$, the costate $p(t)$ is given by $-e^{2(t-T)}$. Then τ is the value of t for which the product of these expressions equals -1.

26.3 Theorem 22.22 does not apply directly, since the dynamics are not autonomous. The time dependence (since it is Lipschitz) can be absorbed into extended dynamics, as follows:

$$x' = y, \quad y' = u/m(z), \quad z' = 1.$$

Note that time t is identified with the extra state component z. With this reformulation, Theorem 22.22 is applicable.

26.4 If the optimal control u_* is neither identically 1 nor identically -1, then the necessary conditions provided by Theorem 22.13 hold normally, and lead to the conclusion.

26.5
(a) One may apply Theorem 23.13.

(b) Show that the maximum principle as given by Theorem 22.22 applies, and must hold in normal form.

(e) Note that this is consistent with the necessary conditions of Theorem 22.22.

26.6 The existence of a solution follows from Theorem 23.11. Expressed in terms of the function $y(t) = x(t) - 2t + 1$, the problem becomes

$$\min \; -\int_0^1 |y(t)|\,dt \;:\; y'(t) \in [-3,-1],\; y(0) = 1,\; y(1) = -1.$$

Apply the extended maximum principle (Theorem 22.26); constancy of the Hamiltonian gives $|y| + \max\{-3p, -p\} = h$. Analysis of the (y, p) phase plane reveals the solution y, and hence the optimal x, which is given by

$$x'(t) = \begin{cases} 1 & \text{if } t \in (0,1/4) \cup (3/4,1) \\ -1 & \text{if } t \in (1/4,3/4). \end{cases}$$

26.7 For (d): Show that the variable-time maximum principle holds in normal form. Deduce from it that the costate is C^∞, as is the time-optimal control.

26.8 Existence theory implies that an optimal process (x_*, u_*, v_*) exists, so the deductive method can be used. Note that the running cost is nondifferentiable. The extended maximum principle can be applied in normal mode. It follows from the adjoint inclusion that p is nonincreasing, with values in $[0,1)$ and $p(0) = 0$. The maximum condition implies that $(u_*, v_*) = (1 - x_*, x_*)$ whenever $p > 0$. The possibility that $p \equiv 0$ on an interval $[T - \varepsilon, T]$ can be ruled out, by using the fact that

$$M(x_*, p) = h = p(1 - x_*) + x_* = x_*(T) \;\; \forall t \in [0,T].$$

Thus $x_*' = 1 - x_*$, and the optimal state trajectory is $x_*(t) = 1 - e^{-t}$. The nature of the solution is different for longer horizons T; see [13, §3.3] for the full analysis of a more complex problem of this type.

26.9 The function $B^* p(t)$ is analytic and not identically zero, and therefore its zeros are isolated.

26.10 As regards existence, Theorem 23.11 cannot be invoked directly, since the dynamics function f does not have linear growth (on the face of it). However, all relevant state trajectories are implicitly constrained to $[0, \bar{x}]$, so f can be redefined… or the direct method could be used, of course. The existence of the turnpike can be deduced from arguments similar to those in § 22.2. See [10] for a considerably harder problem in this vein.

26.12 For (a), use the implicit function theorem.

26.13

(a) From (1) we get $\varphi'(t) < \delta\varphi$; with $\varphi(T) = 0$, deduce $\varphi(t) > 0$ on $[0,T]$. Now use (3) to see that $u'\varphi_u > 0$; conclude with the help of (2).

(b) Use (3) to derive

$$u'\varphi_u = s'(x(t)) - \delta \int_t^T s'(x(\tau))\, e^{-\delta(\tau - t)}\, d\tau.$$

Bearing in mind that s' is nondecreasing, show that the right side is negative.

(c) Prove that $\varphi(t) < 0$ for $t < T$, then examine $u'\varphi_u$. Show that $u' > 0$ initially, and that $u' < 0$ near T.

26.19 For $y \in \mathbb{R}$, let F_+ be the multifunction defined as follows:

$$F_+(x,y) = \{(f(x,u), \omega) : u \in U\},$$

and let $\varphi_+(x,y) = \varphi(x) + y$. With the help of Subbotin's theorem, show that the couple (F_+, φ_+) is weakly decreasing, and exploit this fact.

26.21 For the first assertion, use Exer. 26.19. For the second: if not, show that there exists $p \neq 0$ and $\delta > 0$ such that

$$\langle p, f(x,u)\rangle < 0 \;\; \forall x \in B(0,\delta),\, u \in U.$$

Use this to contradict the preceding conclusion.

26.22 We have seen that (a) implies (b), and that (c) implies (a). It suffices to prove that (b) implies (c). It follows from the compactness of trajectories that $T(\cdot)$ is lower semicontinuous, so it suffices to show that $T(\cdot)$ is upper semicontinuous. Fix any $\alpha_0 \in \mathbb{R}^n \backslash \{0\}$. Given $\varepsilon > 0$, in view of (b), there exists $\Delta > 0$ such that

$$T(y) \leqslant \varepsilon \;\; \forall y \in B(0, \Delta).$$

Now let (x,u) be a minimal-time process from α_0 to 0. By classical results concerning continuity with respect to the initial condition (see [28]), there exists $\delta > 0$ such that

$$\alpha \in B(\alpha_0, \delta) \implies x_\alpha\big(T(\alpha_0)\big) \in B(0, \Delta),$$

where x_α is the state trajectory on the interval $[0, T(\alpha_0)]$ corresponding to the same control u, but with initial condition $x_\alpha(0) = \alpha$. It follows that

$$\alpha \in B(\alpha_0, \delta) \implies T(\alpha) \leqslant T(\alpha_0) + \varepsilon,$$

which verifies the upper semicontinuity of $T(\cdot)$.

26.23
(d) Write the proximal subgradient inequality, and apply the extended maximum principle.

(e) Normalize p_i by dividing by $|(\zeta_i, \psi_i)|$, and take appropriate subsequences.

(f) See Exer 13.17.

26.24 In view of Exer. 26.22, it suffices to prove that the minimal-time function $T(\cdot)$ is continuous at 0. Given $\varepsilon > 0$, take $0 < T < \varepsilon$ such that the process $(0,0)$ is normal on $[0,T]$. Then apply Exer. 26.23 to conclude.

26.29
(a) This follows from applying existence theory and the maximum principle; it is clear in the current context that the adjoint equation and transversality condition characterize a unique costate.

(d) If $\zeta \in \partial_L V(0)$, then $\zeta = \lim_i \zeta_i$, where $\zeta_i \in \partial_P V(\alpha_i)$ and $\alpha_i \to 0$. Let (x_i, u_i) belong to $\Sigma(\alpha_i)$, and let p_i be a corresponding costate for (x_i, u_i). Use sequential compactness arguments in order to show that p_i converges to a costate corresponding to some process $(x,u) \in \Sigma(0)$.

(e) Since V is Lipschitz near 0 and $\partial_L V(0)$ is a singleton, it follows that V is differentiable at 0 (see Exer. 11.15), with $\nabla V(0) = -p_{x,u}(0)$.

(f) If ℓ is convex and Λ is strictly convex, and if the system is linear, it follows that the set $\Sigma(0)$ is a singleton.

26.30
(b) We consider (t,x) as an augmented state whose first component satisfies the dynamics $t' = 1$. Observe that the couple (V,F), where $F(t,x) = \{1\} \times f(x,U)$, is both weakly decreasing and strongly increasing on Ω; this gives (6), by the results of Chapter 12.

(c) If φ is another proximal solution, we may consider weak decrease and strong increase in order to show that φ both majorizes and minorizes V (vocabulary under construction). Details are given in [18, Chap. 4].

References

1. R. A. Adams and F. H. Clarke. Gross's logarithmic Sobolev inequality: A simple proof. *American J. Math.*, 101:1265–1269, 1979.
2. J.-P. Aubin. *Optima and Equilibria*. Graduate Texts in Math. 140. Springer-Verlag, Berlin, 1993.
3. J.-P. Aubin and I. Ekeland. *Applied Nonlinear Analysis*. Wiley Interscience, New York, 1984.
4. M. Bardi and I. Capuzzo-Dolcetta. *Optimal control and viscosity solutions of Hamilton-Jacobi-Bellman equations*. Birkhäuser, Boston, 1997.
5. G. A. Bliss. *Calculus of Variations*. Carus Monograph 1. Math. Assoc. of America, 1978.
6. P. Bousquet and F. H. Clarke. Local Lipschitz continuity of solutions to a problem in the calculus of variations. *J. Differential Equations*, 243:489–503, 2007.
7. S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge Univ. Press, New York, 2004.
8. H. Brezis. *Analyse fonctionnelle*. Masson, Paris, 1983.
9. L. Cesari. *Optimization–Theory and Applications*, volume 17 of *Applications of Mathematics*. Springer-Verlag, New York, 1983.
10. C. W. Clark, F. H. Clarke, and G. R. Munro. The optimal exploitation of renewable resource stocks. *Econometrica*, 47:25–47, 1979.
11. F. H. Clarke. *Necessary Conditions for Nonsmooth Problems in Optimal Control and the Calculus of Variations*. Doctoral thesis, University of Washington, 1973. (Thesis director: R. T. Rockafellar).
12. F. H. Clarke. A classical variational principle for periodic Hamiltonian trajectories. *Proceedings of the Amer. Math. Soc.*, 76:186–188, 1979.
13. F. H. Clarke. *Optimization and Nonsmooth Analysis*. Wiley-Interscience, New York, 1983. Republished as vol. 5 of *Classics in Applied Mathematics*, SIAM, 1990.
14. F. H. Clarke. An indirect method in the calculus of variations. *Trans. Amer. Math. Soc.*, 336:655–673, 1993.
15. F. H. Clarke. Continuity of solutions to a basic problem in the calculus of variations. *Annali della Scuola Normale Superiore di Pisa Cl. Sci. (5)*, 4:511–530, 2005.
16. F. H. Clarke. Necessary Conditions in Dynamic Optimization. *Memoirs of the Amer. Math. Soc.*, 173(816), 2005.
17. F. H. Clarke and M. d. R. de Pinho. Optimal control problems with mixed constraints. *SIAM J. Control Optim.*, 48:4500–4524, 2010.
18. F. H. Clarke, Yu. S. Ledyaev, R. J. Stern, and P. R. Wolenski. *Nonsmooth Analysis and Control Theory*. Graduate Texts in Mathematics, vol. 178. Springer-Verlag, New York, 1998.
19. F. H. Clarke and R. B. Vinter. Applications of optimal multiprocesses. *SIAM J. Control Optim.*, 27:1048–1071, 1989.

F. Clarke, *Functional Analysis, Calculus of Variations and Optimal Control*, Graduate Texts in Mathematics 264, DOI 10.1007/978-1-4471-4820-3,

20. R. Deville, G. Godefroy, and V. Zizler. *Smoothness and Renormings in Banach Spaces*. Pitman Monographs 64. Longman, UK, 1993.
21. N. Dunford and J.T. Schwartz. *Linear Operators Part I*. Wiley Interscience, New York, 1967.
22. R. E. Edwards. *Functional Analysis*. Holt, Rinehart and Winston, New York, 1965.
23. G. M. Ewing. *Calculus of Variations with Applications*. Norton and Company, New York, 1969.
24. W. H. Fleming and R. W. Rishel. *Deterministic and Stochastic Optimal Control*. Springer-Verlag, New York, 1975.
25. I. Fonseca and G. Leoni. *Modern Methods in the Calculus of Variations* : L^p *spaces*. Springer, New York, 2010.
26. I. M. Gelfand and S. V. Fomin. *Calculus of Variations*. Prentice-Hall, Englewood Cliffs, N.J., 1963.
27. H. H. Goldstine. *A History of the Calculus of Variations*. Springer-Verlag, New York, 1980.
28. M. R. Hestenes. *Calculus of Variations and Optimal Control Theory*. Wiley, New York, 1966.
29. J.-B. Hiriart-Urruty and C. Lemaréchal. *Fundamentals of Convex Analysis*. Springer, Berlin, 2001.
30. E. B. Lee and L. Markus. *Foundations of Optimal Control Theory*. Wiley, New York, 1967.
31. C. B. Morrey. *Multiple Integrals in the Calculus of Variations*. Springer-Verlag, New York, 1966.
32. R. R. Phelps. *Convex Functions, Monotone Operators, and Differentiability*. Lecture Notes in Math. 1364. Springer, New York, 1989.
33. L. S. Pontryagin, V. G. Boltyanskii, R. V. Gamkrelidze, and E. F. Mischenko. *The Mathematical Theory of Optimal Processes*. Wiley-Interscience, New York, 1962.
34. R. T. Rockafellar. *Convex Analysis*. Princeton University Press, Princeton, 1970.
35. R. T. Rockafellar and R. Wets. *Variational Analysis*. Springer-Verlag, New York, 1998.
36. H. L. Royden. *Real Analysis*. Macmillan, London, 1968.
37. W. Rudin. *Real and Complex Analysis*. McGraw-Hill, New York, 1966.
38. W. Rudin. *Functional Analysis*. McGraw-Hill, New York, 1973.
39. J. L. Troutman. *Variational Calculus with Elementary Convexity*. Springer-Verlag, New York, 1983.
40. R. B. Vinter. *Optimal Control*. Birkhäuser, Boston, 2000.
41. L. C. Young. *Lectures on the Calculus of Variations and Optimal Control Theory*. Saunders, Philadelphia, 1969.

Index

F. Clarke, *Functional Analysis, Calculus of Variations and Optimal Control*, Graduate Texts in Mathematics 264, DOI 10.1007/978-1-4471-4820-3,